15 August 1999

Final Report

NEAR TERM MODEL DEVELOPMENT Part II

Constantijn Panis
Lee Lillard

RAND
1700 Main Street, PO Box 2138
Santa Monica, CA 90407-2138

Preface

In late 1998, the Social Security Administration, Division of Policy Evaluation, launched a major effort to develop a microsimulation model of retirement income in the year 2020. It awarded two contracts to develop components of the model. The first contract was with The Urban Institute to project income from Social Security benefits, private pensions, private savings, and labor force participation after retirement and to study life cycle patterns of labor force participation and earnings. The results of that effort are documented in Toder, Uccello, O' Hare, Favreault, Ratcliffe, and Smith (1999).

The second contract was with RAND to project demographic transitions, ensure that the distribution of outcomes is preserved in the projections, provide guidance on internal and external consistency, and develop a model of retirement income taxation. This document reports on the findings of this second component.

Many people at RAND have contributed to the project. We especially acknowledge Thierry Cottet, Steven Haider and Jacob Klerman for their substantive contributions. Research programming was provided by Roald Euller, Patricia StClair, and Delia Burroughs. Tanya Burton provided research assistance. Paul Steinberg improved the presentation and readability of this report. Jennifer Wiggin assisted in administrative matters.

We gratefully acknowledge helpful comments from members of MINT' s Panel of Experts, particularly Christopher Bone, John Rust, Alan Gustman, and Finis Welch.

This report was prepared by RAND for the Social Security Administration, Office of Research, Evaluation, and Statistics, Division of Policy Evaluation, under Task Order 0440-98-34176, pursuant to Contract Number 0600-96-27335.

Contents

Figures

Tables

Acronyms

AGI	Adjusted Gross Income
AIME	Average Indexed Monthly Earnings
ARIMA	AutoRegressive Integrated Moving Average
CBO	Congressional Budget Office
CPI	Consumer Price Index for Urban Wage Earners and Clerical Workers
DB	Defined Benefit
DC	Defined Contribution
DI	Disability Insurance
DPE	Division of Policy Evaluation
DRA	Divorce Registration Area
EBRI	Employee Benefits Research Institute
FICA	Federal Insurance Contributions Act
FRB	Federal Reserve Board
GAO	General Accounting Office
HI	Hospital Insurance
HRS	Health and Retirement Study
IRA	Individual Retirement Account
MINT	Modeling Income in the Near Term
MRA	Marriage Registration Area
NCHS	National Center for Health Statistics
NIA	National Institute on Aging
OACT	Office of the Chief Actuary
OASDI	Old-Age, Survivors, and Disability Insurance
OASI	Old-Age and Survivors Insurance
ORES	Office Of Research, Evaluation, and Statistics
PSID	Panel Study of Income Dynamics
SAS	Statistical Analysis System
SCF	Survey of Consumer Finances
SIPP	Survey Of Income and Program Participation
SSA	Social Security Administration
VAR	Vector Auto-Regressive

1. Introduction

1.1. Background

The Division of Policy Evaluation (DPE) at the Social Security Administration (SSA) wants to have the capability to evaluate the distributional impact of policy changes affecting the Social Security system. The outcomes of interest for the DPE are the distribution of future retirement income, marital status, and survival. To evaluate such policy changes, the DPE needs to extend and enhance the current Survey of Income and Program Participation (SIPP)/DPE model to simulate a wide variety of Social Security reform proposals that are under consideration (see, e.g., Olsen 1996, Social Security Advisory Council 1997). These include changes to the contribution rate, changes to the benefit formula, to spousal and widowhood benefits, to the retirement age (already increasing by law to age 67), investment of a portion of the OASDI trust funds in stocks, diversion of a fraction of contributions into personal investment accounts with varying degrees of control by the worker, et cetera. Most proposals are in fact combinations of such changes.

To gain this modeling capacity, the SSA issued a two-part task order to develop a microsimulation model. The first part of the task order, which was issued to The Urban Institute, was to produce projections of various income components. The second part of the task order, issued to RAND, was to project marital and survival status, implement an approach to preserve the distribution of initial values in projected outcomes, ensure internal and external consistency, and generate a taxation model.

The combination of the two parts of the task order will give the SSA a model generating projections of various income components, marital status, and mortality for the cohort born in 1931-1960. In addition to the capability to assess the impact of policy reforms, the model will give the SSA the tools to evaluate the consequences of long-term trend scenarios such as more or less favorable trends in incomes, marriage and divorce rates, and mortality. It is not a structural or dynamic model, which means that it will not predict individuals' behavioral responses to policy changes (e.g., concerning retirement timing). For those types of responses, the SSA may formulate assumptions and conduct scenario analyses.

1.2. Objectives and Approach

This report documents the results of RAND' s work on part two of the SSA task order. In particular, it documents the results of the four substantive tasks specified in the task order: (Chapter 2) Task 1: Demographic Projections, which projects marriage, divorce, widowhood, and mortality transitions for 1990-1993 SIPP respondents born in 1926-1965; (Chapter 3) Task 2: Stochastic Elements, which ensures that the distributions of the income and demographic outcome variables are preserved in the projections; (Chapter 4) Task 3: Consistency Checks, which is concerned with both internal and external model consistency, with the former concerned with corrections for correlation among income components that are projected separately, and the latter concerned with consistency of summary statistics from model projections with external macroeconomic or other models that project similar summary statistics; and (Chapter 5) Task 4: Individual Income Tax Model, which is to develop a model that computes individuals' income tax liabilities.

Appendix A documents all SAS programs used to prepare the SIPP data for analysis and to project demographic histories for the simulation sample. Appendix B includes algorithms and reproduces tax forms and schedules which underlie the individual income tax model.

2. Demographic Projections

2.1. Background

The overall objective of Task 1 is to project demographic transitions for 1990, 1991, 1992, and 1993 SIPP respondents born in 1926-65, including marriage, divorce, widowhood, and mortality.[1] Figure 2.1 illustrates the states of interest and the transitions between them.

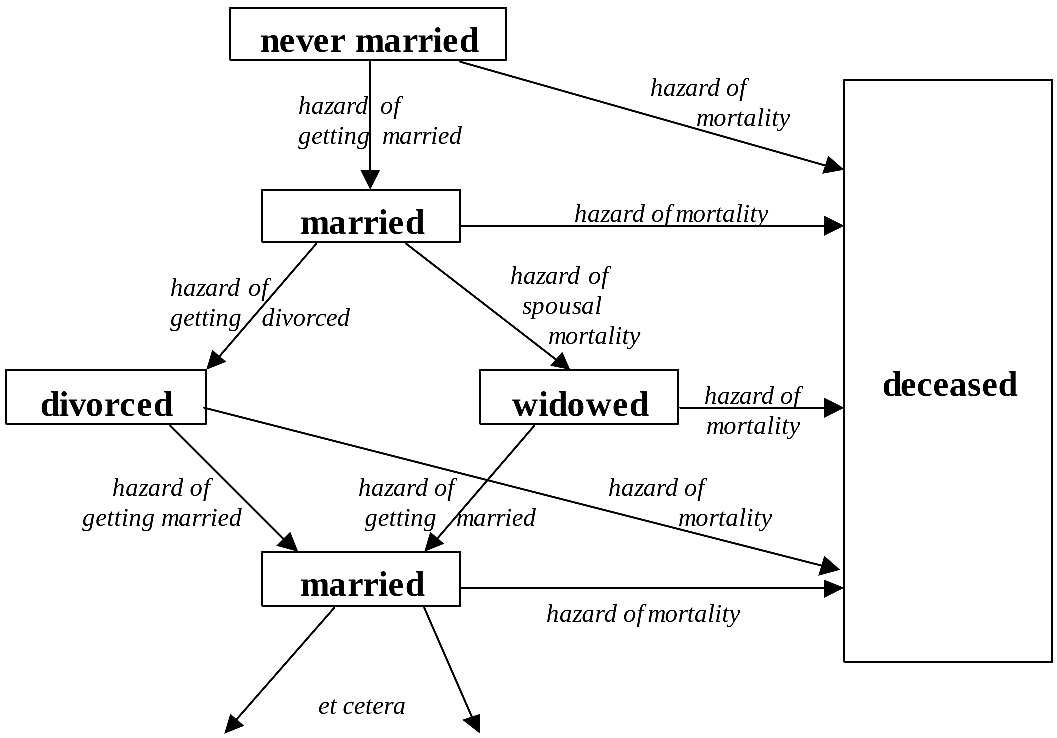

Figure 2.1. Demographic States and Transitions

As shown in Figure 2.1, there are four types of transitions:

- Marriage and remarriage
- Divorce
- Transition into widowhood (spousal mortality)
- Transition to deceased (own mortality)

In addition to marital and survival status, we project the date of onset of disability (not shown in Figure 2.1). Each transition is the outcome of a hazard process, namely the hazards of (re-)marriage, divorce, own and spousal mortality, and onset of disability.

[1] The Statement of Work extends to the 1990 and 1991 SIPP only and restricts the simulation sample to 1931-60 birth cohorts. The projections as described in this document and delivered are a superset of those required by the Statement of Work.

Figure 2.2 illustrates the steps that were involved in producing the demographic projections. First, we estimated model parameter coefficients of the marriage, divorce, mortality, and disability model equations. These models are based on data from 1901-1994 Vital Statistics, the 1968-1994 Panel Study of Income Dynamics, and the 1990 and 1991 waves of the Survey of Income and Program Participation (SIPP). Second, we selected the simulation sample and prepared the data. The simulation sample is based on the 1990, 1991, 1992, and 1993 SIPP waves. Third, we projected respondents' demographic transitions and future states, starting at the last survey date and ending at the time of mortality. These projections take into account the known dates of death between the last survey date and mid-1998, as recorded in SSA's Numident data.

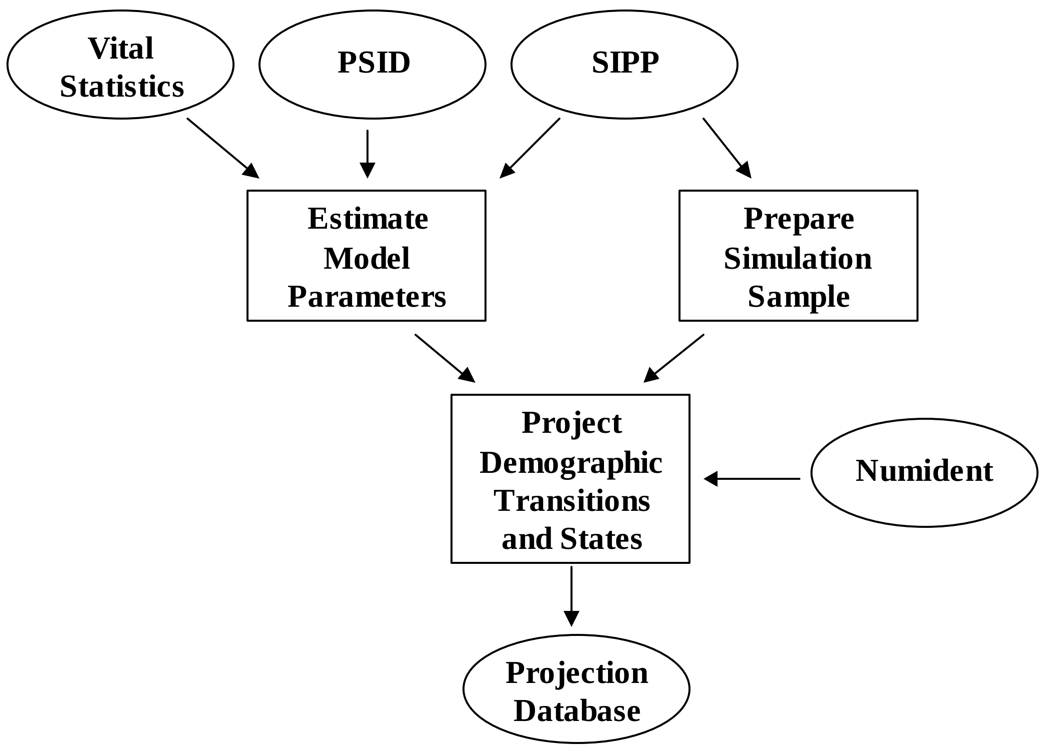

Figure 2.2. Projection Procedure Flow Chart

The sections below describe the estimation procedures and parameter estimates for own and spousal mortality (Section 2.2), marriage and remarriage (Section 2.3), divorce (Section 2.4), and the onset of disability (Section 2.5). Section 2.6 specifies the simulation sample selection criteria and discusses important data preparation issues. Section 2.7 explains the algorithms for projecting demographic states. Section2.8 presents summary statistics of the projections.[2] Appendix A documents the sequence of SAS programs that prepared the data and projected future demographic states.

[2] Chapter 4 compares aggregate projections produced by the MINT model to those produced by other demographic projection models.

2.2. The Model for Mortality

2.2.1. Overview

This section describes estimation of the mortality process parameters. Demographic projections require mortality processes for both the respondent and his or her current and future spouses.

For the current projections, mortality risk is determined by respondents' age, gender, race and ethnicity, educational attainment, permanent household income, and marital status. In addition, the projection method takes account of a secular trend towards increased longevity.

The Survey of Income and Program Participation (SIPP) offers only limited information on dates of death. A relatively small number of respondents dies during the panel period. Dates of death of 1990-93 SIPP respondents between the end of the panel and 1998 are available from administrative records in the Numident file. Mortality specifications that do not involve time-varying covariates may be estimated on these matched administrative records. However, we wish to estimate mortality as a function of time-varying marital status, on which no information is available after the end of the panel. The SIPP/Numident data therefore do not support estimation of our mortality model. Instead we estimate mortality using the Panel Study of Income Dynamics (PSID), a large household survey which has been fielded annually since 1968.[3]

While projections will only be made for respondents born in 1926-65, we estimate mortality models on all cohorts born in or before 1965. The inclusion of older cohorts is important to obtain parameter estimates for elderly persons. The 1926-65 birth cohorts need to be simulated through the year 2020, when the eldest individuals are over 90 years old.

Even though the PSID was designed to be representative of the American population, there may be differences between PSID mortality experiences and those documented in Vital Statistics of the United States. We model such differences (as a function of age, sex, race, and calendar time) and apply a procedure to transform the estimated parameters into parameters that yield projections consistent with Vital Statistics; see below. The resulting mortality hazard parameters are used to project both respondent mortality and spousal mortality (respondent transition into widowhood).

---

[3] We chose the PSID because it has been running for many years, has good information on deaths, marital transitions, and income, and spans the full age range.

2.2.2. The Basic Mortality Pattern

Consider Figure 2.3, which plots the natural logarithm of age-specific mortality rates (log-hazard) for white and black males and females based on 1994 Vital Statistics (National Center for Health Statistics, 1998). Mortality rates decrease sharply during the first twelve years of life, increase during adolescence, stabilize during the early twenties, and increase almost linearly from approximately age 30. The youngest members of our projection sample are around thirty as of the last survey wave, so for our purposes, the baseline log-hazard is almost linear (almost Gompertz). There is some indication in the literature that the mortality log-hazard levels off slightly at higher ages, so we allow for a piecewise linear baseline duration dependency: linear between age 30 and 65, and again linear after age 65.[4]

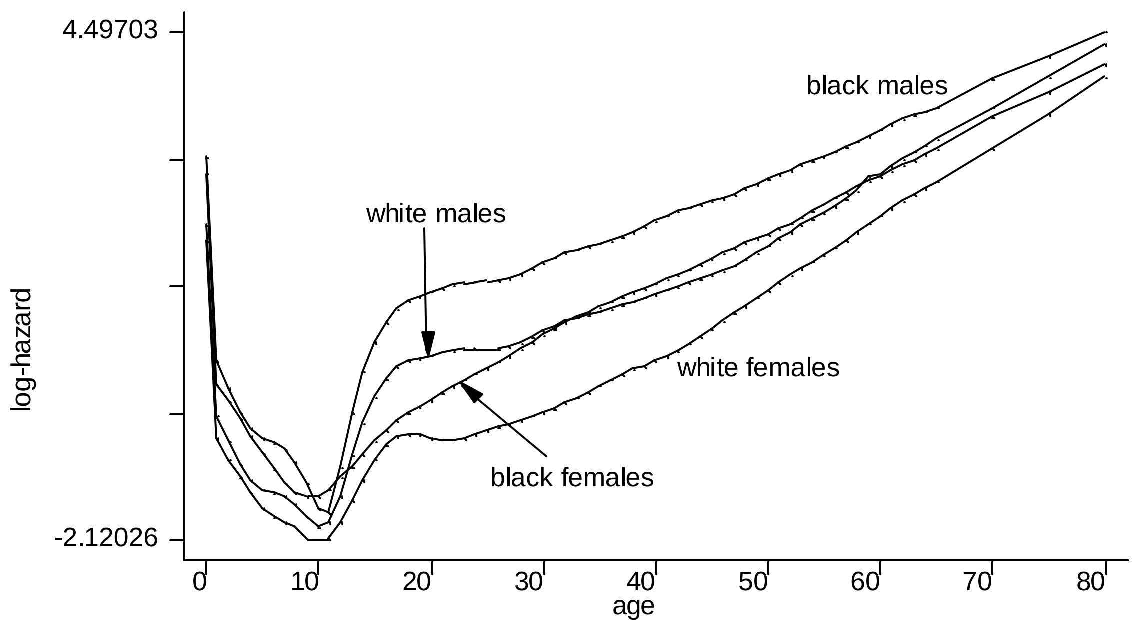

Figure 2.3. Log Death Rates, 1994 Vital Statistics

In line with the literature, we assume that the mortality process follows the standard proportional hazard model (e.g., Kalbfleisch and Prentice 1980):

$$\ln h^{m}(t) = g'T(t) + b'X_{t}, \qquad [2.1]$$

where $\ln h^{m}(t)$ denotes the log-hazard of dying at time t; $g'T(t)$ captures the piecewise-linear age dependency and a linear calendar time trend; and $b'X_{t}$

[4] Various studies suggest a change in the mortality function around age 90, which would call for a change in the slope of the mortality hazard line. Neither the PSID nor Vital Statistics offer sufficient richness to reliable estimate departures above age 90 from the piecewise Gompertz. As noted by Christopher Bone, because of the limited duration of MINT and the cohorts under study, this does not raise any significant issues for this implementation of MINT. The oldest individuals are only in their late eighties by the year 2020, the end of the MINT projection period. It does imply, however, that projections much beyond the year 2020 need to be interpreted with caution.

represents the effects of exogenous covariates: race, educational attainment, marital status, and permanent income.[5] The models are estimated separately for males and females. Measurement of permanent income is described in more detail below. Marital status is a time-varying covariate. Since we are only interested in projecting mortality for individuals who are at least 30 years old, we estimate the model only on PSID respondents age 30 and over. By excluding survival experiences prior to age 30, we avoid the need to carefully account for the irregular log-hazard pattern before age 30, as shown in Figure 2.3. For example, a PSID respondent who was 20 years old as of the first survey in 1968 is included in the estimation sample only starting at his 30th birthday, in 1978 (unless he died or left the sample before 1978, in which case the person does not contribute to the estimates.) Table 2.1 presents the parameter estimates.

Table 2.1. PSID Mortality Hazard Estimates

	Males	Females
Constant	-9.6619 ***	-10.0891 ***
	(.2603)	(.3314)
Age slope 30-65	.0879 ***	.0869 ***
	(.0044)	(.0057)
Age slope 65+	.0793 ***	.0867 ***
	(.0042)	(.0048)
Calendar time	-.0119 ***	-.0152 ***
	(.0038)	(.0047)
Black	.1768 **	.3219 ***
	(.0804)	(.0953)
High school drop-out	.3778 ***	.0934
	(.0704)	(.0778)
College graduate	-.0513	-.2514 *
	(.1040)	(.1427)
Never married	.2138 *	.0184
	(.1132)	(.1421)
Divorced	.4343 ***	-.1185
	(.1146)	(.1527)
Widowed	.1080	-.0041
	(.0905)	(.0805)
Permanent income	-.1591 ***	-.2675 ***
	(.0435)	(.0477)
Income missing	-.4083	-2.1304
	(1.1828)	(3.9271)
Log-Likelihood	-14424.95	

Note: asymptotic standard errors in parentheses;
significance `*' = 10%, `**' = 5%, `***' = 1%

[5] The model of mortality does not control for disability status. The effect of disability, however, is partially captured through our control for permanent income. Also see Subsection 2.9.

All estimated patterns are consistent with well-established findings in the literature. The estimates show that net mortality rates decreased by approximately 1.19 percent (males) and 1.52 percent (females) between 1968 and 1994.[6] Blacks experience significantly higher mortality rates than whites; mortality rates decrease with educational attainment; never married and divorced men face higher mortality rates than married and widowed men, while marital status has almost no effect on women; and mortality risks are lower for individuals with higher incomes.

2.2.3. Differences Between the PSID and Vital Statistics

The PSID was designed to be representative of the American population at the time of its first wave in 1968. Since then, the immigrant composition of the US has changed and there has been some attrition from the PSID. Thus, the PSID may no longer be fully representative of the population. In addition, the PSID interview staff may not be fully successful in recording all deaths, perhaps classifying some deaths as panel attrition. For these reasons, we correct PSID mortality estimates such that they become representative of the American population and so they may be used for projection purposes.

This correction is based on a comparison of PSID mortality and mortality recorded in Vital Statistics of the United States. We collected Vital Statistics data at roughly 10-year intervals between 1901 and 1994 and converted them into mortality hazard spells, similar to the PSID data format. We then estimated mortality hazard models for individuals age 30 and over, using only sex, age, calendar time, and race as determinants. The same specification was run on PSID data. Table 2.2 presents the results. The first column shows estimates based on Vital Statistics; the second on the PSID; and the third, their difference.

Note that estimates based on Vital Statistics have very small standard errors. The reason is that they are weighted by the US population.[7]

Note that we capture mortality reductions over time by a linear trend. SSA's Office of the Chief Actuary (OACT) documents that longevity gains have varied considerably across subperiods of this century. The gains were relatively large between 1968 and 1982, and relatively small between 1982 and 1994 (Bell 1997; see Table 4.4 on page 94). One may debate whether future longevity gains will follow the pace of the entire period since the beginning of this century, or since the establishment of Medicare in the late 1960s, or even since more recent dates. Only time will tell. We take a very long term view and extrapolate from the beginning of this century.

---

[6] The trend in Vital Statistics mortality rates, which will be used for projection purposes, is slightly flatter. See below.

[7] The weights have been divided by 1000, so that standard errors are in fact 1/1000-th of those presented here.

Table 2.2. Differences Between PSID and Vital Statistics

	VS	PSID	PSID-VS
Males			
Constant	-8.3597 ***	-9.6791 ***	-1.3195 ***
	(0.0013)	(0.2480)	(0.2480)
Age slope 30-65	0.0721 ***	0.0909 ***	0.0187 ***
	(0.0000)	(0.0042)	(0.0042)
Age slope 65+	0.0821 ***	0.0838 ***	0.0017
	(0.0000)	(0.0039)	(0.0039)
Time 1901-1994	-0.0081 ***	-0.0179 ***	-0.0099 ***
	(0.0000)	(0.0037)	(0.0037)
Black	0.2815 ***	0.3913 ***	0.1097
	(0.0004)	(0.0778)	(0.0778)
Females			
Constant	-8.7528 ***	-10.2761 ***	-1.5233 ***
	(0.0016)	(0.3260)	(0.3260)
Age slope 30-65	0.0685 ***	0.0902 ***	0.0217 ***
	(0.0000)	(0.0055)	(0.0055)
Age slope 65+	0.0954 ***	0.0862 ***	-0.0093 **
	(0.0000)	(0.0043)	(0.0043)
Time 1901-1994	-0.0141 ***	-0.0181 ***	-0.0040
	(0.0000)	(0.0044)	(0.0044)
Black	0.3325 ***	0.5323 ***	0.1998 **
	(0.0005)	(0.0912)	(0.0912)
Log-Likelihood	-222314824.9	-14498.74	-14498.74

Note: asymptotic standard errors in parentheses;
significance `*' = 10%, `**' = 5%, `***' = 1%

To ensure that our mortality projections, in the aggregate, match those which would be produced by Vital Statistics estimates, we correct the PSID mortality estimates of Table 2.1 by the difference of PSID and Vital Statistics estimates, as in the third column of Table 2.2. *The mortality specification that we use to project dates of death for the SIPP sample is given by Table 2.1 minus the coefficients of the third column of Table 2.2*.

2.2.4. Measurement of Permanent Income

Our measure of permanent income is based on individuals' long-run position in the distribution of household log-income. The SIPP panels contain 32 monthly household income values (eight waves with four monthly values each). SAS program `income.sas` groups these into the first 10 values, the next 12 values, and the last 10 values. Program `perminc.sas` rescales these sums such that they represent annual values and estimates a very simple model in which annual log-income is regressed on age (piecewise linear with different slopes before and after age 50), sex interacted with marital status (never married, divorced, and widowed relative to married), and a measure of number of adult-equivalents in the household.[8] Table 2.3 shows the results of this regression.

For each respondent and each of his or her three annual incomes, we computed the residual. For each respondent, we computed the average of his or her three residuals and took this average as a measure of permanent income. The same procedure was applied to the PSID. While many more than three annual household income measures are available in the PSID, we restricted ourselves to the first three incomes after the respondent reached age 30, so as to be compatible with the SIPP measurement.

Table 2.3. Household Log-Income Parameter Estimates

Constant	9.3733 *** (0.0215)
Age slope 25-50	0.0110 *** (0.0005)
Age slope 50+	-0.0156 *** (0.0004)
Never married male	-0.1267 *** (0.0129)
Never married female	-0.3486 *** (0.0138)
Divorced male	-0.1916 *** (0.0162)
Divorced female	-0.4963 *** (0.0136)
Widowed male	-0.2016 *** (0.0267)
Widowed female	-0.3876 *** (0.0136)
log(adults equivalent)	0.7541 *** (0.0118)

Note: asymptotic standard errors in parentheses; significance `*' = 10%, `**' = 5%, `***' = 1%

As shown in Table 2.1, our measure of permanent income is strongly predictive of mortality risk.[9]

[8] This measure, log(adults+0.7*kids)$^{0.65}$, is based on recent research on poverty measurement, which suggests that (adults+0.7*kids)$^{0.65}$ is a reasonable conversion of adults and children in a household into adult need equivalents.

[9] An alternative measure of permanent income is the respondent's Average Indexed Monthly Earnings (AIME), or an equivalent summary measure computed for younger workers. This measure is available in the SIPP from matched SSA records and may be computed in the PSID from self-reported information. However, years in non-covered employment cannot be distinguished from years with zero earnings in matched SSA records. This is a potentially serious limitation, especially for earlier years when Social Security coverage was far from universal.

Table 2.4 shows remaining life expectancies for a 60-year-old in 1990 by sex, race, and a combination of permanent income and education. This table is generated from parameter estimates of Table 2.1 (corrected by Table 2.2) for stereotypical values of the covariates.[10] The income points correspond to the first quartile, median, and third quartile. Our model controls for both income and education, which are highly correlated. Projections of life expectancies by income, holding education constant, would therefore understate differences by income. We therefore show projections by income, assuming that the lower incomes have less than a high school education, the median are high school graduates, and the third quartile corresponds to college graduates. "Q1 income—high school drop-out" represents a high school drop-out whose permanent income measure is equal to the first quartile cut-off; "Median income—high school graduate" represents a high school graduate with median permanent income; and "Q3 income—college graduate" represents a college graduate with permanent income equal to the third quartile cut-off.

Table 2.4. Remaining Life Expectancies at Age 60 by Sex, Race, and Income/Education

	Male	Female
White		
Q1 income—high school drop-out	16.4	23.9
Median income—high school graduate	20.6	26.1
Q3 income—college graduate	21.8	29.8
Black		
Q1 income—high school drop-out	15.7	22.7
Median income—high school graduate	19.7	25.0
Q3 income—college graduate	20.9	28.6

Note that life expectancy differences between the first and third income quartile cut-offs are between five and six years. This has important implications for poverty in old age. As projected by The Urban Institute/Brookings Institution, individuals with low lifetime income may enter retirement with limited financial resources. As projected by RAND, these resources will need to support a shorter retirement period, on average, than experienced by higher-income and better-educated individuals. It also has important implications for the degree of progressivity that is implicit in the Social Security program (Panis and Lillard, 1996).

[10] The table contains "cohort" life expectancies and may not be directly compared to standard "current" life expectancies as published in Vital Statistics publications; see Section 4.4.1 for the definition.

2.3. The Model for Marriage and Remarriage

In line with the literature, we model the transitions into marriage using a continuous time hazard model, also known as a failure-time model (e.g., Kalbfleisch and Prentice, 1980). Its basic form is given by piecewise-linear Gompertz. The multiplicative effects on covariates on the hazard are equivalent to additive effects on the log-hazard:

$$\ln h_{ij}^{w}(t) = \Gamma_w(t) + q_w' X_{ij} \qquad [2.2]$$

where $\ln h_{ij}^{w}(t)$ is the log-hazard that individual i marries (w for wedding) for the j-th time. The marriage baseline hazard, $\Gamma_w(t)$, captures duration dependencies on respondent age and duration since the previous marriage dissolved. In addition, as discussed below, $\Gamma_w(t)$ may include a duration dependency on calendar time to capture secular changes in marriage rates. All covariates are constant within spells; some, such as the number of previous marriages, differ across marriages, but do not vary over time within a spell. Throughout we suppress the person subscript.

The transition into (re-)marriage involves a period during which the individual is unmarried and "at risk" of marrying. Once married, the individual is no longer at risk of marrying. (We assume monogamy.) Instead, he/she enters a new period in which he/she is at risk of divorcing. Alternatively, the marriage may end through the death of the person's spouse. After the divorce or widowhood, the individual enters a new period in which he/she is at risk of re-marrying. The marriage and remarriage processes are thus naturally captured by hazard models, also known as failure-time models.

We do not account for unobserved heterogeneity, even though it has been shown to be significant in our own earlier work and not independent of mortality risk (Lillard and Panis, 1998b). The reasons for this exclusion here are that the projection exercise would be very much more complicated (and thus impossible to complete within the required time frame) and that it would rely on distributional assumptions that would undoubtedly be controversial. To our knowledge, no one has worked out the technique for projections of hazard processes that are based on random effects heterogeneity. For purposes of the Near Term Model, exclusion of heterogeneity is not a severe limitation. The main purpose of the Model is to yield accurate predictions, not to estimate structural parameters with behavioral interpretations. A model without heterogeneity but with extensive controls for parity (marriage number) will generate accurate predictions. We experimented extensively with parity controls, both in additive and interactive form.

2.3.1. The Data

The model may be estimated on any data set that contains longitudinal information on marriage and divorce. The SIPP itself is an excellent candidate, as is the PSID with which we have ample experience (Lillard 1993; Lillard and Panis 1996, 1998a, 1998b; Panis and Lillard 1996).[11] Since the SIPP population is the population on which projections will be based, we propose to estimate models of marriage on the SIPP panels. Only the 1990 and 1991 SIPP panels are used for estimating the (re-) marriage process; the 1992 and 1993 panels are used to assess out-of-sample goodness-of-fit; see Subsection 2.3.2 (page 27).

Marital History Data Quality Issues

Although SIPP data files are among the cleanest of all major longitudinal surveys, some data quality issues inevitably arise. We highlight the most important marriage history issues.

SIPP marriage history information is only obtained for the first two and the most recent marriage. If respondents were married more than three times, we do not know how many times exactly, or the dates when they married, divorced, and/or widowed. We imputed the number of marriages and transition types/dates based on the PSID, which contains full information. We estimated a simple ordered probit model of number of marriages, using the period between the dissolution of the second marriage and the most recent wedding date as sole explanatory variable. (No other variable was found to be predictive.) We then stochastically imputed the number of SIPP marriages based on the same gap measure. Dissolution types (divorce versus widowhood) were randomly assigned based on the fractions found to divorce (85.1 percent) and end in widowhood (14.9 percent) in the PSID. Transition dates were selected such that marriages were spread evenly between the dissolution of the second marriage and the wedding of the most recent marriage.[12]

Marriage transition dates are reported to the month only. Very short marriages and very short divorce/widowhood spells were therefore sometimes reported to result in multiple transition dates in the same month. Instead of selecting the 15-th of the month as our best-guess transition date, we chose the 10-th and the 20-th for the two dates.

We updated marriage histories as reported in the Wave 2 Topical Modules with panel information through the end of the survey sequence. In quite a few cases, the status reported for month 9 was not the same as in the Topical module. In many cases, a legitimate transition was the most likely cause. The remaining cases followed the basic rule that the marital status as of the last marriage described by the topical module was correct. The monthly series was adjusted accordingly starting in month 9 to be consistent with the last observed marital status on the topical module. Processing forward from month 9 to 32/36/40 (depending on the number of SIPP waves), we recorded any changes in marital status. The details of this consistency

[11] Our prior work focused on the timing of marital separation rather than divorce.

[12] For estimation purposes, windows were created around best-guess transition dates that were as wide as possible, so that the additional marriages contribute through their parity but very little through their timing.

adjustment are extensively documented in the source code itself (`updatemar.sas`). There were cases that transitioned from never married to divorced or widowed or from separated to married; in each case, a general rule was formulated to resolve the issue as well as possible.

In a handful of cases, respondents reported a first marriage date before their birth date. In a few dozen cases, first marriages presumably took place before age 12. We accepted such respondents' reports in the sense that we took them as baseline for the projections, but we did not use them to estimate models of getting (re-)married and divorced.

All inconsistencies were flagged by assigning non-zero values to variable `marqual`. Only "clean" marriage histories were used in estimating hazard models of getting (re-)married and divorced.

Explanatory Covariates

As is well known from the literature (including our own contributions), age, sex, education, and race/ethnicity are powerful predictors of marital status changes. In addition, the timing of a remarriage is determined by the duration since the previous marriage ended and the current marital status (divorced or widowed); the timing of a divorce is determined by the duration since the wedding. All these factors will be incorporated in the duration dependencies $\Gamma_w(t)$ and the covariates X_{ij}.

We did not control for spousal compatibility measures such as the difference in age between husband and wife, differences in race/ethnicity, and the difference in educational attainment. Spousal characteristics are not available for marriages that were completed prior to the first SIPP interview and may thus not be used for estimation purposes.

Another powerful predictor of marital transitions is the number of children that the couple has (and the number born outside marriage or brought in from prior marriages). The main problem with such measures is that their values are unknown for the projection period. One would need to develop additional models for fertility (separately for marital and nonmarital), and project future births. The issue is further complicated by evidence that childbearing is endogenous to divorce risk (Lillard 1993), so that systems of simultaneous hazard equations with correlated heterogeneity would need to be developed, estimated, and projected. This would be a huge undertaking, well beyond the scope of the current project and with only very small benefits to the current project.[13]

[13] As noted by reviewer John Rust, the policy applicability of MINT would be greatly enhanced if MINT were expanded to a closed overlapping generations model of the full U.S. population. To that end, the value of including a fertility module would be very high.

There has been a marked trend in marriage rates in the United States. Figure 2.4 shows the number of marriages per 1,000 unmarried women age 15 and over for 1940 through 1990 (NCHS 1995a) and indicates a steady decline in the marriage rate from 1947 to the current time. We control for a linear time trend in our marriage model specification to capture changes over time not accounted for by other covariates in the model.

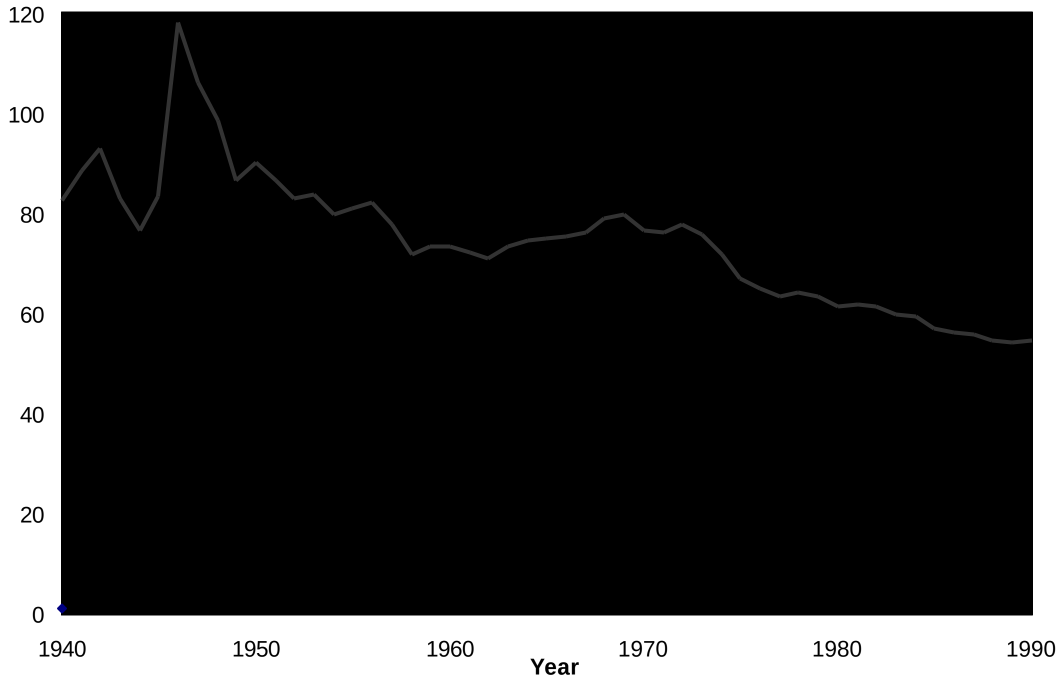

Figure 2.4. Marriages per 1,000 Unmarried Women Aged 15+, 1940-1990

Table 2.5 shows parameter estimates of the marriage and remarriage process, estimated separately for males and females. The age pattern indicates that marriage rates increase until age 20 and decrease thereafter. For remarriage, the hazard increases during the first three years after dissolution of the previous marriage, and decrease thereafter. Marriage rates are decreasing over time, consistent with Figure 2.4. The hazard of marriage for the second, third, and subsequent times are higher than for the first time. (This may be due to heterogeneity rather than marriage number; see above.) White non-Hispanic persons are more likely to enter a marriage than other races and ethnicities. Men who are high school drop-outs tend to marry later than high school graduates, whereas the pattern is reversed for women. College graduates tend to marry later than high school graduates. Men with a high permanent income, measured as explained in subsection 2.2.4 (page 19), tend to marry sooner; their female counterparts later.

Table 2.5. Estimates of Marriage Formation

	Males	Females
Constant	-23.7332 ***	-21.9557 ***
	(1.2834)	(.5813)
Age slope 0-16	1.1847 ***	1.1783 ***
	(.0813)	(.0370)
Age slope 16-20	.6211 ***	.3855 ***
	(.0121)	(.0072)
Age slope 20-25	.0840 ***	-.0545 ***
	(.0041)	(.0038)
Age slope 25+	-.0496 ***	-.0751 ***
	(.0010)	(.0012)
Slope on duration unmarried,	.1208 ***	.0789 ***
0-3 years	(.0153)	(.0146)
Slope on duration unmarried,	-.1086 ***	-.0726 ***
3-8 years	(.0101)	(.0094)
Slope on duration unmarried,	-.0382 ***	-.0223 ***
8+ years	(.0074)	(.0061)
Calendar time	-.0079 ***	-.0036 ***
	(.0004)	(.0003)
Married once before	.4325 ***	.3590 ***
	(.0327)	(.0304)
Married twice before	.6669 ***	.6248 ***
	(.0425)	(.0395)
Married three or more times before	1.2981 ***	1.2017 ***
	(.0576)	(.0506)
Black	-.3587 ***	-.5179 ***
	(.0208)	(.0183)
American Indian, Eskimo or Aleut	-.1756 **	-.0543
	(.0750)	(.0647)
Asian or Pacific Islander	-.2368 ***	-.2276 ***
	(.0491)	(.0425)
Hispanic	-.0592 **	-.3009 ***
	(.0241)	(.0232)
High school drop-out	-.0744 ***	.1284 ***
	(.0153)	(.0134)
College graduate	-.1733 ***	-.4313 ***
	(.0153)	(.0173)
Widowed	.2856 ***	-.3813 ***
	(.0399)	(.0356)
Permanent income	.0164 ***	-.0279 ***
	(.0059)	(.0049)
Log-Likelihood	-328,842.85	

Note: asymptotic standard errors in parentheses;
significance `*' = 10%, `**' = 5%, `***' = 1%

2.3.2. Goodness of Fit of Marriage Transition Models

Our hazard models of getting married and divorced (Section 2.4, below) are based on experiences of the 1990 and 1991 SIPP respondents. We applied these estimates to 1992 and 1993 SIPP respondents to assess the goodness of fit. Starting all respondents at age 12 (when no one is married yet), we projected marital transitions until the last interview date. Table 2.6 shows actual marital status and projected marital status for these 1992 and 1993 respondents.

Table 2.6. Actual and Projected Marital Status

	Actual status		Projected status	
	Frequency	Percent	Frequency	Percent
Never married	4301	11.3	3954	10.4
Married	28065	73.7	27198	71.4
Widowed	1382	3.6	1863	4.9
Divorced	4346	11.4	5079	13.3

As is clear from the table, projected and actual marital status distributions are very close. The discrepancies may be due to stochasticity (because of duration draws in the projection method) or to a mild self-selection. The projection namely assumes that all respondents survive through the last survey. In reality, SIPP respondents are the survivors of their birth cohorts, and thus somewhat self-selected.

The distributions of projected number of marriages and age at first marriage are also very close to the actual distributions (not shown here; see `checkmar.sas`).

2.4. The Model for Divorce

Similar to the model for marriage formation, we model marriage dissolutions using a continuous time hazard model:

$$\ln h_{ij}^{d}(t) = \Gamma_d(t) + q_d' X_{ij} \qquad [2.3]$$

where $\ln h_{ij}^{d}(t)$ is the log-hazard of divorcing (d) for the j-th time. The baseline hazard, $\Gamma_d(t)$, captures duration dependencies on the duration since the wedding and respondent age. In addition, as discussed below, $\Gamma_d(t)$ includes a duration dependency on calendar time to capture secular changes in divorce rates. All covariates are constant over the duration of the divorce spell; some, such as the number of previous marriages, differ across marriages, but do not vary over time within a marriage. For reasons discussed above, we do not account for unobserved heterogeneity. Throughout, we suppress the person subscript.

Marriages that end in widowhood will never result in a divorce. These marriages thus contribute censored dissolution spells. Similarly, marriages that are still in progress at the last interview date contribute a censored spell. Hazard models offer a natural way to incorporate such censored durations.

As with the model for marriage and remarriage, we use 1990 and 1991 SIPP data to estimate our model of divorce behavior. Data from the 1992 and 1993 panels were used to assess goodness-of-fit; see above.

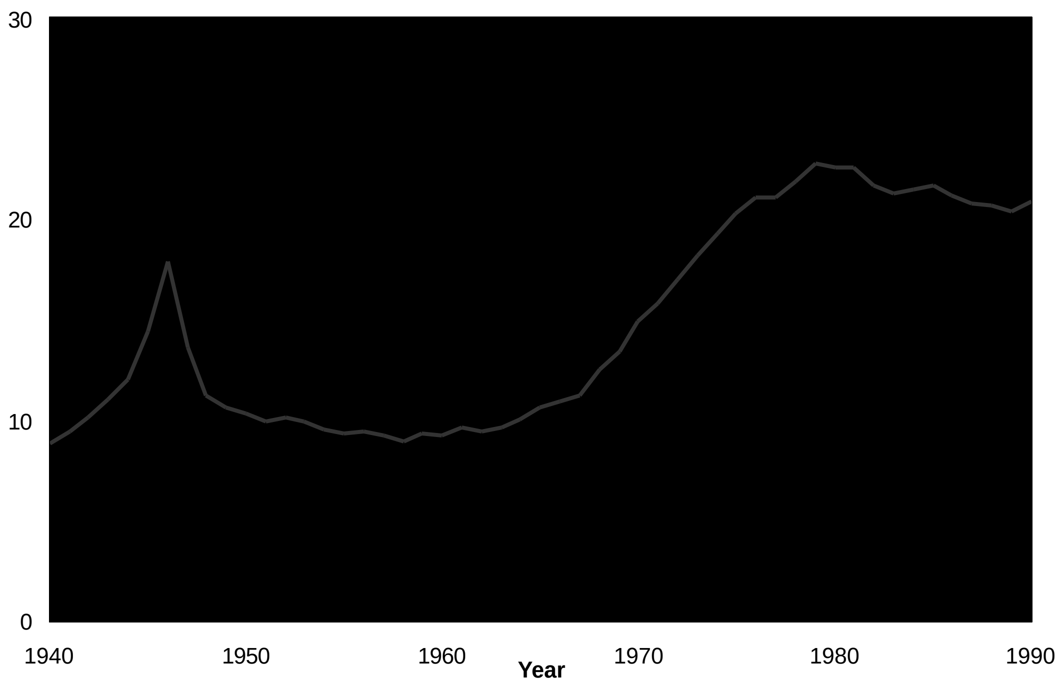

Figure 2.5. Divorce Rate per 1,000 Married Women Aged 15+, 1940-1990

Figure 2.5 shows the divorce rate per 1,000 married women aged 15 and over in the United States from 1940 to 1990 (NCHS 1995b). The divorce rate increased steadily between 1960 and 1980; since 1980, the trend has been approximately flat. We therefore include a piecewise-linear time trend in our divorce specification with a node at 1980. Table 2.7 shows the parameter estimates.

Table 2.7. Divorce Hazard Estimates

	Male	Female
Constant	-1.0198***	-1.7268***
	(.1100)	(.0946)
Age slope, 0-30 years	-.1193***	-.1021***
	(.0038)	(.0032)
Age slope, 30+ years	-.0400***	-.0523***
	(.0015)	(.0015)
Marriage duration, 0-1 years	.4339***	.7350***
	(.0724)	(.0694)
Marriage duration, 1-4 years	.2395***	.1526***
	(.0117)	(.0107)
Marriage duration, 4-15 years	-.0228***	-.0156***
	(.0032)	(.0030)
Marriage duration, 15-25 years	-.0386***	-.0275***
	(.0048)	(.0044)
Marriage duration, 25+ years	-.0875***	-.0832***
	(.0060)	(.0052)
Calendar time, pre-1980	.0401***	.0429***
	(.0010)	(.0008)
Calendar time, post-1980	-.0025	.0058***
	(.0020)	(.0019)
Second marriage	.5737***	.6368***
	(.0248)	(.0232)
Third or higher marriage	1.2503***	1.3584***
	(.0396)	(.0338)
High school drop-out	-.0274	-.0085
	(.0208)	(.0186)
College graduate	-.2117***	-.1068***
	(.0204)	(.0215)
Black	.1198***	.1786***
	(.0276)	(.0240)
American Indian, Eskimo or Aleut	.3339***	.3237***
	(.0766)	(.0611)
Asian or Pacific Islander	-.6198***	-.6378***
	(.0692)	(.0610)
Hispanic	-.3015***	-.2076***
	(.0343)	(.0314)
Log-Likelihood	-687,975.70	

Note: asymptotic standard errors in parentheses;
significance `*' = 10%, `**' = 5%, `***' = 1%

Table 2.7 indicates that divorce rates decrease with age. They increase during the first four years of marriage and decline as the marriage lasts longer. The estimate of the time trend parameters confirms the trend in Figure 2.5: divorce rates increased significantly until 1980 and remained almost unchanged since then. (Our projection algorithms assume that the post-1980 trend continues to the year 2020.) Divorce rates are higher for second and subsequent marriages than for first marriages. (This may be due to heterogeneity rather than marriage number; see above.) Blacks and native Americans experience higher divorce rates than whites, Asians, and Pacific Islanders. Hispanics experience lower divorce rates than non-Hispanics.

Given that husbands and wives always get divorced at the same time, we would ideally want to estimate the divorce equation at the couple level, i.e., controlling for both spouses' characteristics including spousal compatibility measures. However, spousal characteristics are only known for marriages that were ongoing during the SIPP panel. The characteristics of former spouses are unknown. It is therefore impossible to estimate the divorce equation at the couple level.

2.5. The Model for Onset of Disability

Our demographic projections do not involve health or disability status. However, The Urban Institute and Brookings Institution found self-reported functional disability to be strongly predictive of earnings. In addition to marital and survival status, we therefore project disability status.

Disability is defined as self-reported functional disability: "Does ... have a physical, mental, or other health condition which limits the kind or amount of work ... can do?" We simplify reality by assuming that disability is an absorbing state, i.e., one cannot recover. We model the timing of onset of first disability report.

The SIPP, in its Work Disability History Topical Module, asks for functional disability. If functional disability is present, the date of onset is asked. The Work Disability History Topical Module is only administered to respondents age 16-67. Our model for the onset of disability is based on pooled observations from the 1990 and 1991 SIPP. Since the main objective is to project dates of disability onset for individuals at least around 30 years of age, we only include respondents in the estimation data set that are not disabled as of their 30^{th} birthday. In other words, the disability spells upon which our estimates are based all begin at age 30 and continue through either the date of disability onset or the interview date. Respondents that indicated being disabled but who did not provide a date of onset were excluded from the estimation sample.

Table 2.8 presents the results of estimation. The risk of becoming disabled increases with age and accelerates after one's 45^{th} birthday. There is no significant difference between males and females. High school drop-outs are far more likely to become disabled than high school graduates; college graduate experience even lower disability rates. Asians and Pacific Islanders face the lowest disability risks, followed by whites and blacks. Native Americans experience the highest disability rates. Individuals of Hispanic origin are less likely to become disabled than non-Hispanics.

Table 2.8. Estimates of Onset of Disability

Constant	-7.3766 ***
	(0.1786)
Age slope, 30-45	0.0526 ***
	(0.0045)
Age slope, 45+	0.1746 ***
	(0.0047)
Male	0.0062
	(0.0348)
High school drop-out	0.7312 ***
	(0.0389)
College graduate	-0.6668 ***
	(0.0577)
Black	0.2779 ***
	(0.0487)
American Indian, Eskimo or Aleut	0.5446 ***
	(0.1465)
Asian or Pacific Islander	-0.5249 ***
	(0.1378)
Hispanic	-0.1674 **
	(0.0681)
Log-Likelihood	-25736.61

Note: asymptotic standard errors in parentheses;
significance `*' = 10%, `**' = 5%, `***' = 1%

2.6. Data Preparation Issues

Most of the data preparation was applied to all respondents to the 1990, 1991, 1992, and 1993 SIPP panels, regardless of birth year, so that models of (re-)marriage, divorce, and disability could be estimated on all age ranges. However, demographic transitions are projected only for respondents born in 1926-65 (boundaries inclusive). The sample selection criteria are:

1. Year of birth not before 1926 and not after 1965; AND
2. Strictly positive value for full-panel person weight (`pnlwgt`) OR be present until the last interview wave.

Extensive exploration of the data revealed some puzzling issues related to person weight. First, variable `pnlwgt` is zero for approximately 15 percent of individuals present in all interview waves. Second, `pnlwgt` is often nonzero for individuals who left the sample before the full panel was administered. Third, `pnlwgt` is very frequently nonzero for 1992 panel respondents who only participated in nine of the ten 1992 interviews.

Consultation with SIPP experts Denton Vaughan and Judy Eargle indicated the following.[14] Nonzero `pnlwgt` values for individuals who did not respond to all interviews may be legitimate where the individual was deceased, entered an institution, moved into military barracks, moved abroad, or otherwise became ineligible for follow-up. The frequent occurrence of nonzero `pnlwgt` for 1992 respondents who participated in all but the last interview is explained by the government shutdown of December 1994 which forced the Census Bureau to cancel follow-up interviews with at least one rotation group. Zero `pnlwgt` values for about 15 percent of individuals who participated in all interviews remain a mystery. Regardless of the exact explanations, *the Census Bureau recommends that policy analysis should be based on cases with strictly positive* `pnlwgt` *only.*

Table 2.9. Simulation Sample Sizes

	`pnlwgt`		
	0	>0	Total
Birth year 1926-30	1,952	6,656	8,608
Birth year 1931-60	24,960	59,537	84,497
Birth year 1961-65	7,998	11,968	19,966
Total	34,910	78,161	113,071

Table 2.9 shows the number of observations in the simulation sample.[15] The total sample size is 113,071. Of these, 34,910 have a zero value of `pnlwgt` even though

[14] E-mail communication from Denton Vaughan to Howard Iams of January 5, 1999.

[15] The 1993 SIPP panel contains two male respondents that reported being married: IDs 7451101.11.101 and 7451101.11.102. Given the need to project the same potential divorce date for

they participated until the last interview. *The sample for analysis purposes, with strictly positive `pnlwgt` values, consists of 78,161 individuals.* Of these, 59,537 are born in 1931-60 (the cohorts that were specified in the Scope of Work); an additional 18,624 are born during the five years on either side of the 1931-60 birth cohorts.

Marital Status Issues

As discussed above, there were some issues with the quality of marital status reports. We updated marriage histories as reported in the Wave 2 Topical Modules with panel information through the end of the survey sequence. In quite a few cases, the status reported for month 9 was not the same as in the Topical Module; in other cases, respondents went through "illegal" transitions (never married to divorced, separated to married, etc.) from one month to the next. In a handful of cases, respondents reported a first marriage date before their birth date. In a few dozen cases, first marriages presumably took place before age 12.

For purposes of estimating models of marriage and divorce, we dropped individuals with poor reporting quality. For purposes of projecting future transitions and states, however, we included all respondents regardless of the quality of their reports. As a general rule, we assume that the most recent marital status report is correct, so that the projections (which start at the last interview date) are based on the most recent marital status report. The projection data set (`mint.sd2`) contains both historical and future marital transitions; the historical transitions reflect our best judgment of actual transitions.

Disability Status Issues

The SIPP records disability status and date of onset in the Work Disability History Topical Module. This module is only administered to respondents age 16-67. All MINT simulation respondents are in that age range and should have been administered the Topical Module. However, disability status is missing for 16,921 respondents in the projection sample. In addition, disability status is unknown for former spouses, i.e., individuals to whom the respondent was married, but either deceased or divorced before the SIPP panel. Furthermore, 1,697 individuals reported being disabled but did not provide a date of onset. We imputed disability status and/or onset date for these groups.

The imputation algorithms are identical to those being used for future projections of disability status and other hazard outcomes; see Section 2.7 below. For former spouses and for respondents with missing disability status, we imputed an onset date. If that date fell before the interview date, we coded variable `disabled`=1 (to indicate

---

both spouses, a special algorithm would need to be developed to ensure spousal consistency for this couple. Instead, we dropped this couple from the simulation data.

that he or she became disable before death) and `disabdte` to indicate the date of onset, as imputed. If the imputed date was after the last interview, we coded `disabled`=0 and left the onset date, `disabdte`, equal to missing. For projection purposes, these individuals were treated identically to respondents who indicated that they were not disabled. For the 1,697 individuals that reported being disabled, but did not provide an onset date, we imputed an onset date under the restriction that the date fall before the interview date.

Spousal Characteristics

Projections of widowhood and divorce dates require information about both own and spousal characteristics: spousal sex, date of birth, race, ethnicity, and education. In addition, for The Urban Institute/Brookings Institution to project future earnings, disability status and the date of disability onset are required. These characteristics are only known for spouses who were themselves respondents to the SIPP surveys. Even if they participated in only one interview, we recorded their characteristics.

By request of The Urban Institute/Brookings Institution, we imputed spousal characteristics for former spouses. The imputation algorithms are based on empirical couple distributions in the SIPP data. Consider imputations of race. We cross-tabulated the races of husbands and wives in the data. To assign the race of a former spouse of, say, an Asian person, we drew a uniformly distributed random number and assigned a race according to the empirical distribution of spousal races: spouses of Asian persons were white, black, Asian, and Native American in 52, 0, 46, and 2 percent of the cases, respectively. The race of a former spouse of a white, black, or Native American person was imputed in a similar manner. Similarly, a Hispanic person marries another Hispanic person in 87 percent of the cases; a non-Hispanic person marries another non-Hispanic person about 99 percent of the cases. Similarly, the education of a high school graduate's former spouse was assigned based on the finding that high school graduates marry high school drop-outs in 12 percent of the cases, high school graduates in 72 percent of the cases, and college graduates in 16 percent of the cases. Spousal dates of birth were imputed using the empirical distribution of the age difference between husbands and wives. Imputations of spousal disability status and date of onset of disability were based on the disability model, as discussed above, and not on the empirical joint distribution of husbands' and wives' disability statuses.

Spousal characteristics are recorded in array variables. For example, educational attainment of spouses are recorded in variables `speduc1` through `speduc8`, allowing for up to eight spouses (marriages) per respondent.

2.7. Projection Algorithms

Figure 2.1 (page 13) shows potential demographic transitions and the hazard processes that drive their timing. The projection method is as follows. As of the last interview wave, an individual finds himself in any one of the demographic states shown in Figure 2.1. Depending on the state, he is subject to two or more transition hazards. For example, suppose the person is never married. He may become (1) married or (2) deceased. His next state is determined by whichever transition takes place first. To this end, we generate two durations, namely until marriage and death. The various demographic states affect each others' transition hazards, but conditional on observables, all hazard processes are statistically uncorrelated, and we may generate durations independently.

The probability that a generic event has not happened yet as of time t is by definition given by its survivor function, *S(t)*, i.e., by one minus the cumulative probability function, 1-*F(t)*. The hazard is by definition the relative decline of the survivor function,

$$h(t) = -\frac{dS(t)/dt}{S(t)}, \qquad [2.4]$$

so that, still by definition,

$$S(t) = \exp\left\{-\int_{t=t_0}^{t} h(t)dt\right\}, \qquad [2.5]$$

where t_0 is the time at which the event became at risk of occurrence. The median duration t^m until an event occurs is given by the solution to $S(t^m) = \frac{1}{2}$, i.e., $t^m = S^{-1}(\frac{1}{2})$. Note that all hazard models in our projection exercise are of the general form

$$\ln h(t) = gT(t) + b'X_t, \qquad [2.6]$$

i.e., the log-hazard is piecewise-linear in durations *t*. This implies that there is a closed-form solution to the survivor function and also to its inverse, i.e., the expected duration may be found by a closed form computation. In addition to being very flexible, piecewise-linear duration dependencies have the advantage that all computations have a closed-form solution, i.e., no numerical integration is required.

For purposes of projecting dates of death and other demographic transitions, the expected duration is not the desired concept, since it would lead to predictions that all respondents die exactly after their remaining life expectancy. (Also see Chapter 3 on Stochastic Elements.) Instead, we draw randomly from the distribution of durations. This is accomplished by drawing a random number between 0 and 1, say, S^*, and solving for the duration t^* as

$$S^* = S(t^*) \quad \Leftrightarrow \quad t^* = S^{-1}(S^*). \qquad [2.7]$$

For each potential transition, we draw a duration. The shortest duration determines which transition occurs. For example, the duration until marriage for a never married person may be 5 years, while the duration until death may be 30 years. We conclude that the person will marry first. He now becomes subject to the competing hazards of divorce, widowhood, and death. Note that the mortality hazard is a function of marital status; now that the person has married, he faces more favorable survival chances. We therefore draw a new duration until death, taking account of the married status. In addition, durations are drawn until divorce and until the spouse' s date of death. Whichever of the three randomly drawn durations comes first determines the next transition. This process continues until the person becomes deceased.[16]

2.7.1. Information from Numident Files

The last interview waves of the 1990 and 1991 SIPP panels took place sometime in 1992-94. Between then and June 1998, we know with certainty from administrative Numident records that some individuals deceased. The same issue arises with 1992 and 1993 SIPP respondents, whose Numident records are up-to-date through October 1998. Unfortunately, the Numident file is not complete, i.e., some individuals may have become deceased without corresponding record in the Numident file. Consider Figure 2.6, which graphs the natural logarithms of mortality rates based on Numident information (with 1990 and 1991 SIPP respondents as denominator) and 1994 Vital Statistics.

As is clear from visual inspection, death rates from Numident records are lower than they should be according to Vital Statistics. In other words, we can project deaths from Numident records with certainty, but we must generate additional deaths between the last survey wave and June 1998, when the Numident records were created. To this end, SAS program file `match.sas`generates mortality projections through June 1, 1998 (for 1990-91 SIPP respondents), and October 1, 1998 (1992-93 SIPP respondents). It finds that 2.6 percent of the sample (1,869 individuals) should be deceased as of 6/1/1998 or 10/1/1998, but that Numident records only show a death rate of 2.0 percent (1,444 individuals). In other words, the Numident records only appear to cover 77 percent (1,444/1,869) of the SIPP population.

[16] An alternative approach would project respondents' life paths in discrete steps, such as months. For example, the probability that a never married man marries during the next month is *p*; if this probability exceeds a randomly drawn variable (from a uniform distribution between 0 and 1), we project the wedding to occur. Similarly, project whether a death occurs and select the dominant transition. Then repeat for each subsequent month until death. This approach may be more suitable in models which incorporate covariates that vary frequently with time. It has limitations where multiple transitions are projected without a clearly dominant one, such as divorce and widowhood.

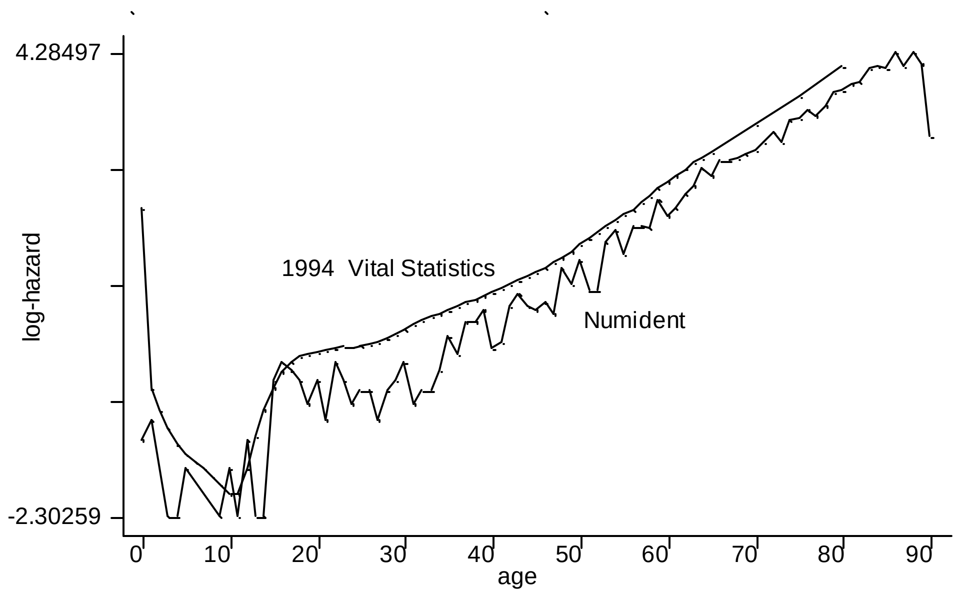

Figure 2.6. Log Death Rates, Numident vs 1994 Vital Statistics

We therefore randomly assign 77 percent of the SIPP sample as "matched" and 23 percent as non-matched. (This is random, except that the 1,444 deceased individuals are matched with certainty.) The projection method in `mint.sas` then distinguishes three types of individuals:

- 1,444 individuals are deceased with certainty in the month indicated by the Numident records.
- The remaining of the 77 percent "matched" respondents are guaranteed to survive through 6/1/1998 or 10/1/1998; after that date, the program accepts randomly generated survival durations.
- For the 23 percent non-matched respondents, the normal duration projections apply at all dates after the last interview wave.

It should be noted that the Numident records are subject to imperfect data quality. SSA staff matched the SIPP surveys to SSA's Master Numident file, and provided RAND with four small Numident files, for 1990, 1991, 1992, and 1993 SIPP respondents. These files contained a total of 10,228 records, each representing one deceased respondent. In many cases, the Numident month of death occurred before the final interview wave. Most of these indicated that the respondent died shortly before the last interview. We accepted the Numident information as correct provided that the Numident date was three months or less before the last interview. In 727 cases, the Numident date occurred more than three months before the last interview date; we assumed that an error was made in SSA's matching procedure and ignored Numident information for these respondents. In 39 cases, the Numident ID could not be matched to a SIPP individual. Again, we assumed that incorrect cases were pulled from the Master Numident file, and ignored Numident information for these 39 cases.

In six cases, there were duplicate Numident records. We randomly selected one Numident record from each pair and ignored the other information.

2.7.2. Characteristics of Future Spouses

Every time a respondent is projected to marry or remarry, a new spouse needs to be taken into consideration. Newly entering spouses do not become observations themselves; they only appear as new elements of variable arrays with spousal characteristics. The simulation database contains characteristics of every spouse, whether they are relevant before, during, or after the SIPP interviews. The characteristics include spousal sex, date of birth, race, ethnicity, education, disability status and date of disability onset. These characteristics are only known with certainty for spouses who were themselves respondents to the SIPP surveys. Even if they participated in only one interview, we recorded their characteristics. Above we explained how we imputed characteristics of former spouses. Characteristics of future spouses are imputed in exactly the same manner. They are directly relevant for the projection of spousal death dates (widowhood dates).

2.7.3. Spousal Consistency

Our projection algorithms are designed to ensure spousal consistency: if a couple is married as of the last interview date, we project the next transition to be on the same date for husband and wife. If the first transition is a divorce, we project the same divorce date for husband and wife; if the first transition is the husband's death, we project that the wife becomes widowed on that same date; and similarly, we project his widowhood date to be at her death date.

Spousal consistency is achieved by using the same random number seed for husbands and wives. Three potential transitions are relevant: divorce, his death, and her death. His death only involves his characteristics, i.e., respondent characteristics when processing his projections, and spousal characteristics when projecting her future. The mirror case arises for her date of death, i.e., his widowhood date. Projections of the divorce date generate an additional complexity because divorce equations are estimated separately for males and females (Table 2.7, page 17).[17] If we were to use respondents' own characteristics, different divorce dates would be generated, even if the seeds were equal. In light of indications that women's marriage history reports tend to be of higher quality than men's, we project divorce dates based on the wife's characteristics (and the female divorce model coefficients), if available. In other words, projections of a divorce date of a woman are always based on her own characteristics and the female divorce model. Projections of a divorce date of a man

[17] It is impossible to estimate divorce equations using both the husband's and the wife's characteristics, including measures of spousal compatibility such as whether they are of the same race, because spousal characteristics are only known for marriages that are still ongoing at the time of the SIPP interviews. Characteristics of former spouses are imputed and thus contain only noise.

are based on his wife's characteristics if these are known with certainty, i.e., for marriages that were in progress at the last interview date. Divorce dates of men's future marriages are based on his own characteristics (and the male divorce specification).

2.8. Projection Results

The previous section explained how we project individuals' life course, starting at the last interview date and ending at the date of death. We generate variables and variable arrays to record all transitions: array `marb` for marriage begin dates; array `mare` for marriage end dates; array `howend` for the type of marriage disposition (divorce, widowhood, own death); variable `disabled` for whether the person became disabled before death; variable `disabdte` for the onset of disability (if `disabled`=1); and variable `deathdte` for the date of death.

SAS macro `%figstat` may be used to determine individual' s demographic status at any particular date. The following tables show the projected demographic distribution as of January 1, 2020. Table 2.10 tabulates demographic status for all 1990-93 SIPP respondents born in 1931-60; Table 2.11 is conditional on survival through 2020.

Table 2.10. Projected Demographic Distribution in 2020
(in percent; 1931-60 birth cohort)

	Male	Female	Total
Never married	3.5	5.0	4.2
Married	47.3	39.7	43.4
Widowed	3.7	19.6	11.8
Divorced	8.2	14.9	11.6
Deceased	37.4	20.9	29.0
Total	100.0	100.0	100.0

Table 2.11. Projected Demographic Distribution in 2020
(in percent; 1931-60 birth cohort, survivors only)

	Male	Female	Total
Never married	5.5	6.3	6.0
Married	75.6	50.2	61.2
Widowed	5.8	24.8	16.6
Divorced	13.1	18.8	16.3
Total	100.0	100.0	100.0

We project that 29 percent of the respondents in the simulation sample will be deceased by the year 2020. Almost twice as many men as women will have become deceased. Among the survivors, 6 percent will have never married, 61 percent will be married, whereas the remaining one-third will be equally divided between divorced and widowed. However, there will be many fewer widowers than widows. Only about 6 percent of surviving men will be widowed, compared with one out of four surviving women.

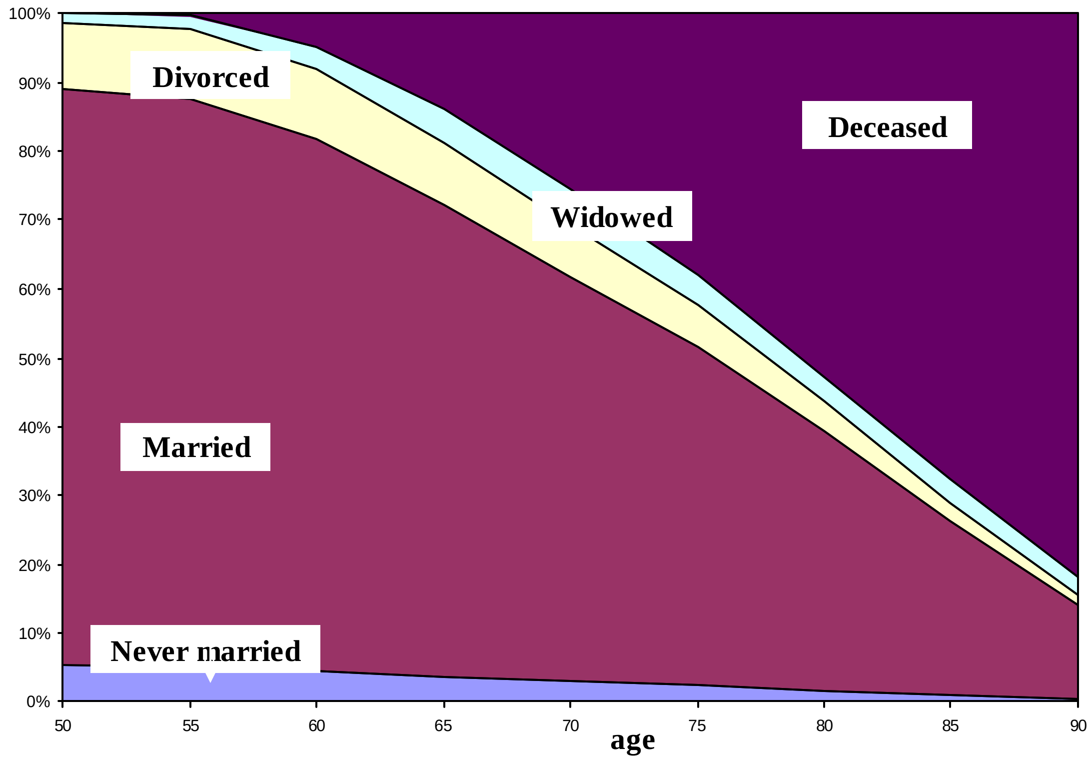

Figure 2.7. Life Cycle Composition: Men Born in 1931-40

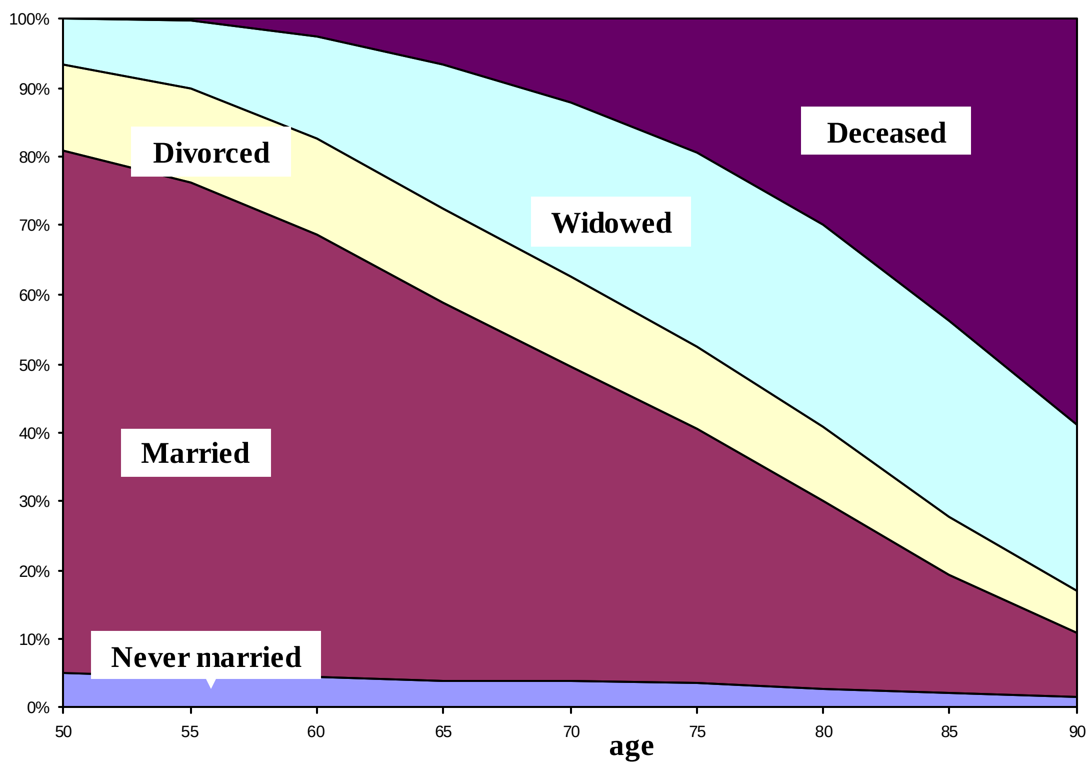

Figure 2.8. Life Cycle Composition: Women Born in 1931-40

Figure 2.7 and Figure 2.8 show the life cycle composition of men and women born in 1931-40, respectively. The figures start at age 50, when the youngest in this cohort participate in SIPP interviews. As of age 50, all are thus alive to participate in SIPP surveys. The upper bound is age 90, corresponding to the year 2021 for the oldest individuals. As individuals age, an increasing number becomes deceased or widowed. Note the large differences between men and women: men remain overwhelmingly married, but as their numbers become smaller, a large fraction of women becomes widowed.

Figure 2.9 and Figure 2.10 show life cycle demographic compositions of surviving men and women born in 1931-1940, i.e., similar to the previous two figures but without the deceased category. The distribution of men by marital status remains virtually unchanged, with predominantly (re)married men. Women, on the other hand, become increasingly widowed at advanced ages.

An important reason for developing the MINT microsimulation model, as opposed to a macro model, is the ability to determine program eligibility for individuals based on individuals' unique characteristics. Consider Figure 2.10, which shows that about 14.5 percent of women born in 1931-40 reach age 62 as divorcées. What fraction of these women will be able to claim Social Security benefits on the basis of their ex-husbands' earnings? Of all individuals that reach age 62 as in divorced status, Table 2.12 shows the fraction whose most recent marriage lasted less than ten years. Overall, about 39 percent of divorced women reach age 62 without a claim on spousal benefits. (In addition, 43 percent of divorced men cannot claim spousal benefits, but they are more likely to have had substantial earnings themselves.) The ineligible fraction is increasing by birth cohort. *Overall, 3.2 million divorced women in the 1931-60 cohort will not be eligible for spousal benefits.* To determine how many of these women would have had sufficiently low lifetime earnings so as to collect spousal benefits, one needs to consider the earnings projections as produced by The Urban Institute/Brookings Institution.

Table 2.12. Fraction Divorced Individuals Married Less Than Ten Years

	Male	Female	Total
1931-40 cohort	40.9	31.1	34.9
1941-50 cohort	42.9	39.9	41.1
1951-60 cohort	44.1	41.4	42.5
Total	43.1	38.9	40.6

A similar calculation may be carried out to determine the fraction of widows ineligible for widowhood benefits because they were married less than nine months (Social Security Handbook §401). MINT projects that 30.2 million women in the 1931-60 birth cohorts become widowed. Of these, about 220,000 (0.7 percent) were married less than nine months. In addition, about 40,000 men became widowed less than nine months after their wedding.

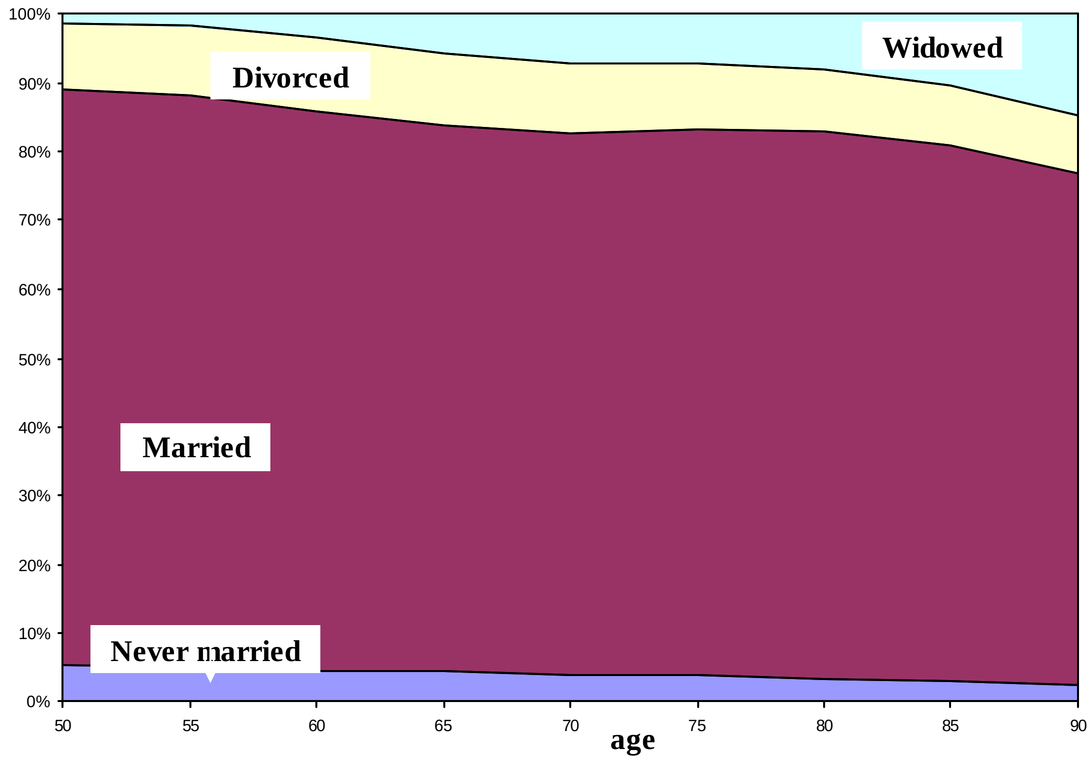

Figure 2.9. Life Cycle Composition: Surviving Men Born in 1931-40

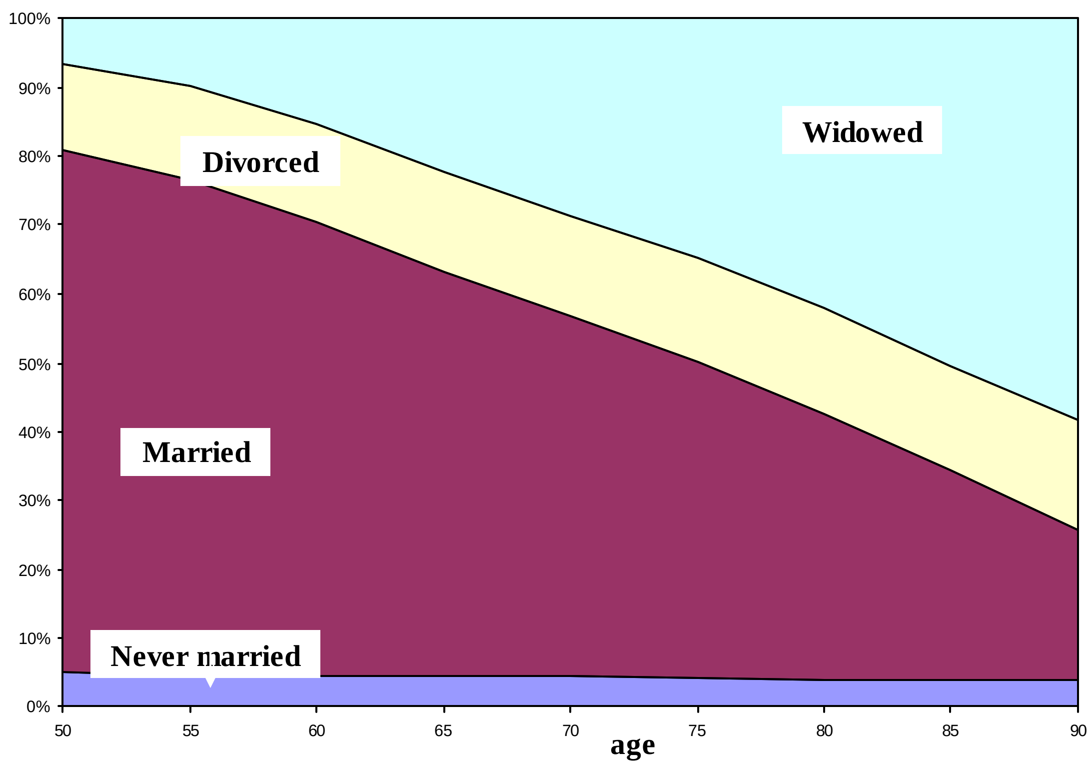

Figure 2.10. Life Cycle Composition: Surviving Women Born in 1931-40

Chapter 4 returns to our demographic projections and compares them to those produced by SSA's Office of the Chief Actuary.

2.9. Mortality as a Function of Disability Status

As documented above, survival projections are based on a mortality hazard model which does not account for disability status. However, disability status is a strong predictor of survival, as shown in Table 2.13. Disabled males face mortality risks that are 2.45 (=exp(0.8971)) times as high as those experienced by their disability-free counterparts, whereas disability increased women' s mortality risk by a factor of 2.94 (=exp(1.0816)).[18]

Note that the effect of permanent income on mortality risk is substantially smaller than in the specification without control for disability status (Table 2.1).

The projection algorithms support longevity projections which take account of disability status. See Appendix A.7 for an explanation of how to modify the projection program such that the projections are based on the specification with account of disability status of Table 2.13.[19]

[18] The PSID did not collect disability status of wives in earlier waves. Married women that became deceased early in the panel thus often have missing disability status, which explains the positive and significant coefficient on missing disability status for women.

[19] It should be noted that account for disability status requires more than modification of the longevity projection algorithms. In particular, income from assets as projected by The Urban Institute assumes that families purchase an lifelong joint and survivor annuity with 80 percent of their assets. The current annuitization algorithms do not take account of disability status of husbands or wives. They need to be modified for consistency throughout all MINT components, such that the disability-free face less generous annuity tables than the disabled.

Table 2.13. PSID Mortality Hazard Estimates, Controlling for Disability Status

	Males	Females
Constant	-9.4700 ***	-10.0894 ***
	(0.2626)	(0.3428)
Age slope 30-65	0.0758 ***	0.0774 ***
	(0.0046)	(0.0058)
Age slope 65+	0.0679 ***	0.0787 ***
	(0.0043)	(0.0048)
Calendar time	-0.0156 ***	-0.0229 ***
	(0.0039)	(0.0051)
Black	0.1810 **	0.2560 ***
	(0.0811)	(0.0962)
High school drop-out	0.3350 ***	0.0515
	(0.0710)	(0.0778)
College graduate	0.0139	-0.2422 *
	(0.1032)	(0.1404)
Never married	0.2725 **	-0.0609
	(0.1130)	(0.1442)
Divorced	0.3525 ***	-0.2248
	(0.1134)	(0.1564)
Widowed	0.1305	-0.1530 *
	(0.0922)	(0.0843)
Disabled	0.8971 ***	1.0816 ***
	(0.0779)	(0.1032)
Disability status missing	0.0815	0.5158 ***
	(0.3591)	(0.1387)
Permanent income	-0.0851 *	-0.1920 ***
	(0.0441)	(0.0499)
Income missing	-0.4138	-1.9734
	(1.1782)	(3.9264)
Log-Likelihood	-14287.14	

Note: asymptotic standard errors in parentheses;
significance `*' = 10%, `**' = 5%, `***' = 1%

3. Stochastic Elements

3.1. Introduction and Summary

This chapter contains the contents of the three Letter Reports that we produced in support of Subtask 2, "Stochastic Elements." This subtask is concerned with a technical issue, namely to ensure that the diversities of demographic and economic outcomes at baseline are preserved in the projections. We identified areas in model components of both The Urban Institute/Brookings Institution and RAND where special care was required to preserve such diversities. Our recommendations have subsequently been incorporated in projections of The Urban Institute/Brookings Institution and RAND. As a result, this chapter is no longer necessary to understand MINT. Indeed, the reader may wish to skip to Chapter 4 (page 71). Below we largely replicate reports that were produced under this subtask, without editing the contents in any substantively important way. They reflect an evolving discussion of the relevant issues.

Before replicating the reports, we briefly summarize the issue and our recommendations. As detailed below, variation across individuals in the outcomes of interest is created by many factors. The "stochastic elements" referred to in the Statement of Work are economic and fiscal variables that affect the income and demographic outcomes of interest. However, residual variation is at least of equal importance to ensure that the distribution of outcomes is preserved in projections.

The first letter report discussed the inclusion of residual variation into projections based on a variety of functional forms. That discussion is replicated below as Section 3.2.4. It is our understanding that The Urban Institute/Brookings Institution have added residual variation to all its projections. Similarly, the RAND demographic projections are all based on random draws from duration distributions and thus incorporate hazard models' implicit residual variation.

The main economic and fiscal variables for which we recommend stochastic variation are (1) returns on equity and bonds (mostly relevant for defined contribution plan balances) and (2) employer match rates of defined contribution pension plan contributions. It is our understanding that The Urban Institute/Brookings Institution have incorporated such variation into their projections.

The second letter report, included as Section 3.3 below, developed three alternatives for incorporating stochastic elements. The first option is to draw once from the distribution; the second option is to project the outcome of interest under multiple (Monte Carlo) draws of the stochastic variables and compute the average; and the third option is to replicate the data many times and draw accordingly many values. We recommended the first option for its simplicity, except in cases where the sample size of the subpopulation of interest is small. Few such sample sizes appear to be small, especially after the MINT contractors agreed to extend the projections to 1992 and 1993 SIPP panels. The Urban Institute/Brookings Institution followed our recommendation and drew single values for the stochastic variables.

The third letter report, included as Section 3.4 below, outlined the ways in which MINT incorporates stochastic variation. Prompted by comments from Christopher Bone, it also developed some theoretical considerations for determining the smallest subpopulation for which MINT is capable of generating reliable distributional consequences. There is no single solution to this issue; it depends on the "Type I Error" (related to significance levels) and "Type II Error" (related to power of the predictions) that one is willing to tolerate. As pointed out by John Rust, however, it is always possible to reduce idiosyncratic noise from single replications by running the model simulations sufficiently many times, until additional replications no longer change the outcome(s) of interest by more than a user-specified tolerance level.

3.2. Specify Stochastic Elements of the Model

3.2.1. Objectives

The overall objective of Task 2 is to ensure that the distributions of the income and demographic outcome variables are preserved in the projections. The goal of this subtask is to identify the economic or fiscal variables that should be stochastic in the model, recommend appropriate distributions for them, and identify logical linkages between stochastic elements needed to maintain internal consistency at the person or family level.

3.2.2. Overview

At this stage of the overall project, many details on model structures and implementation strategies are not yet known. It is therefore not always possible to identify exactly where the simulation model would benefit from the implementation of stochasticity in future values of economic or fiscal variables. In the discussion below, we use our best judgment on the structures of the various submodels based on currently available letter reports and the presentations and discussions during the meeting with the Panel of Experts of June 3, 1998. We make several implicit assumptions and recommend alternative scenarios for cases where the model structure appeared insufficiently well-defined.

Variation across individuals in the outcomes of interest is created by many factors. The focus of this section, in accordance with the RFTOP, is on economic and fiscal variables that affect the income and demographic outcomes of interest. In addition, behavioral differences across individuals, such as in savings rates and in job turnover, may also often be important. We will assume that such behavioral variation is either ignored or accounted for in the modeling strategy. We will return to this issue toward the end of this section and illustrate how even very crude models of individual behavior generate additional variation.

A major source of variation in the forecast is uncertainty about the future values of the variables or assumed parameters used to make the forecast. These values may pertain to regressors, to externally determined parameters that are imposed on the projection model (such as the rate of return on assets), or to other assumptions (such as the assumption that all employers match 50 percent of their employees' DC pension plan contributions). Projections that ignore variation in such future values lose part of the resulting variation in the outcomes of interest.

However, in light of the objective of preserving the current population distributions in the projections, a second source of variation is critically important—perhaps even more important than economic and fiscal variables. No model is likely to fit the observed data perfectly, and projections based on expected values are necessarily subject to (at least) the same error rates. One source of variation in the outcomes around the forecast value is therefore the variability not explained by the underlying model. In linear regression models, such as those used to model income components, such *residual variation* is explicitly part of the model. In qualitative outcome models, such as those used in both income and demographic transitions, residual variation may be explicit or implicit in projected (participation/survival) probabilities.

The RFTOP only refers to variation in future values of covariates or assumed parameters. In the next section, we identify economic or fiscal variables that should be stochastic in the model, recommend appropriate distributions for them, and identify logical linkages needed to maintain internal consistency. Section 3.2.4 develops approaches to adding residual variation to the generic types of models the Urban Institute and RAND are likely to apply. Given that these approaches require access to residuals as predicted from the estimation procedure, and in light of the fact that the RFTOP did not refer to residual variation as part of Task 2, we propose that the Urban Institute and RAND incorporate residual variation in their respective projections. We would appreciate an explicit statement on this issue from the Task Manager.

3.2.3. Economic or Fiscal Stochastic Elements

We now discuss stochastic elements of an economic or fiscal nature. The richness of the submodels suggests many candidates for model elements that vary across individuals or households. In this section, we highlight the factors we deem most important as measured by their potential impact on the distribution of projected outcomes. We discuss these factors by Task, to the extent relevant, so as to facilitate subsequent implementation of an approach to incorporating stochasticity.

Part I, Task 2: Project retirement income from assets and savings

This task is concerned with projecting retirement income from non-pension assets and savings. The two predominant determinants of such income are the individual (or household) saving rate and the rate of return received on assets and savings.

The individual saving rate is a behavioral factor that is part of the Urban Institute model. As outlined in John O'Hare' s letter report on this subtask of May 29, 1998, the model will distinguish various saver types (life-cycle savers, precautionary motive savers, and bequest motive savers). Presumably, the model will incorporate variation in savings rates and/or savings patterns across saver types. The implementation is not yet specified, so it is unclear whether the model allows for variation in saving rates within saver types, as present in the population. Urban' s model development will provide insight in the magnitude of this variation; it may be that incorporating additional stochastic variation does little to account for the distribution of asset income. Since saving rates are behavioral, not economic or fiscal, we will not discuss them in any detail.

The rate of return is an average of returns on various asset categories -- such as stocks, bonds, pass book savings, cash holdings, real estate, own business assets, etc. -- weighted by individuals' portfolio composition shares. Since individuals' portfolios differ, the rate of return varies across individuals. Furthermore, rates of returns on a single asset type may vary over time, generating additional variation.

The Urban Institute model considers both total wealth and homeowner equity. It is not yet clear to us exactly what role the rate of return on total wealth (termed the "rate of appreciation") will play in the model. Will it be estimated as part of the model or imposed from an outside source? It is also not clear yet what role homeowner equity will play in the model. It appears that homeowner equity is part of the total wealth measure, so that the rate of return is an overall rate of return.

The rate of appreciation, *r*, is not subscripted by individual or calendar time in O'Hare's model. The assumption that the rate is constant across individuals and over time, however, is probably overly stringent. *We propose that the projections be subject to rates of return that vary both across individuals and over time.*

We now discuss several issues that arise in determining an appropriate distribution for the rate of return.

- Variation over time in the rate of return on various asset categories may be determined from publications such as Ibbotson Associates (1998) and O'Shaughnessy (1998). See, for example, Table 3.1, taken from Ibbotson Associates (1998).

Table 3.1. Summary Statistics of Annual Returns

	Geometric Mean	Arithmetic Mean	Standard Deviation
Large company stocks	11.0%	13.0%	20.3%
Small company stocks	12.7	17.7	33.9
Long-term corporate bonds	5.7	6.1	8.7
Long-term government Bonds	5.2	5.6	9.2
Intermediate-term government	5.3	5.4	5.7
U.S. Treasury Bills	3.8	3.8	3.2
Inflation	3.1	3.2	4.5

These data are compiled from year-by-year total returns on each of the asset categories. Those year-by-year total returns mask, however, that there is substantial variation in the rate of return of assets *within* each category. For example, the large company stocks data are currently based on the Composite Index of Standard & Poor's 500 largest American companies. However, individual investors rarely hold a portfolio that reflects the relative shares in the S&P 500, so that the rates of return experienced by individual investors vary, even in a single year. The standard deviations presented in Table 3.1 are thus underestimates of actual variation.

Publications such as *The Wall Street Journal* may provide data to compute the standard deviation of the rate of return on individual stocks or bonds. Given that most individuals have some level of diversification in their portfolio holdings, the standard deviation of returns on individual stocks and bonds yield an upper bound on the risks experienced by individuals.

To our knowledge, no data source permits a detailed investigation of portfolio shares on individual investors down to individual assets of all types. We therefore need to make a plausible assumption on the rate of return on total assets within broad asset classes.

- Apart from the issue of pooling of assets into major categories, additional variation arises because individuals differ in their relative portfolio shares. There are several surveys that permit an analysis of asset share holdings. Among them are the SIPP, the Survey of Consumer Finances (SCF), and the PSID. A fruitful approach would be to determine the distribution of individuals' portfolio allocations into broad categories as dictated by the survey(s), possibly conditional on observables such as age. In the MINT projection stage, one would draw from the portfolio allocation distribution and from the distributions of rates of return within asset class to obtain a rate of return for an individual or household.

- A final issue relates to the relevant accumulation period or time horizon. The rates of return affect asset accumulation over a long period of time, both before and after retirement. The compound rate of return has a smaller variance than year-by-year rates of return. It may, however, not be necessary to determine the distribution of compound rates of return over various accumulation periods. An approach whereby draws are applied from year-by-year rates of return will result in compound yields with a smaller variance. If there is a strong autoregressive component to rates of return, it would be missed by this approach. A priori, it is our judgment that the costs of an extensive exploration of this issue outweigh the benefits in terms of preserving the distribution of income from assets and savings.

Part I, Task 3: Project retirement income from pensions

For this task, it is important to distinguish defined benefit (DB) and defined contribution (DC) plans.

An important determinant of the income flows from DB plans is job tenure and thus job turnover. As of today, it is not clear whether and how job transitions will be incorporated in the Urban Institute model. Absence of such a model in favor of some fixed-rule assumption would lose much of the heterogeneity that exists in the population, especially for more recent cohorts who are younger at the last survey date. This behavioral factor is likely to generate much of the variation we observe in receipt of DB pensions benefits.

It should be noted that many retirees receive income from both DB and DC plans and that their entitlement may have accrued at various points over their career. This implies that, for example, workers who at the end of the 1990-1991 SIPP panels are only covered under a DC plan may well switch to a job with DB plan coverage during the projection period. The assumption that workers remain on their current job, or any model that does not allow for workers switching between DB and DC plan coverage, would miss much of the variation in both DB and DC pension receipt.[20]

For DC pensions, three determinants with variation across individuals play a particularly important role.

- Employee contribution rate. Employees vary widely in their contribution rates, mostly within the upper legal limit. This behavioral factor will presumably be part of the Urban Institute model.

- Employer contribution match rate. As discussed by Cori Uccello during the meeting with the Panel of Experts of June 3, 1998, there is variation across employers in their contribution match rate. There appears to be an inverse relationship between the employee contribution rate and the employer match rate. A priori, it is not clear whether variation in employer match rates has substantial implications for variation in income from DC plans. *We recommend that the extent of variation be explored and, if it is substantial, that stochastic variation be added to the employer match rate.* The appropriate distribution may be determined from an analysis of IRS Form 5500 data or from the database that the Employee Benefits Research Institute (EBRI) recently collected. This database contains 401(k) plan information for 23,000 plans with over 2.5 million participants and $75 billion of assets. EBRI may be willing to share some of these data or carry out an analysis for the benefit of this project, as indicated by Jack VanderHei in the meeting of June 3, 1998. Alternatively, GAO (1996) and EBRI/Greenwald (1995) offer guidance on this issue.

- The rate of return on plan assets. The issues here are similar to those raised under Task I-2, retirement income from assets and savings. An important difference is that DC plan assets tend to be more diversified into mutual equity and bond funds than non-pension assets. The rate of return on various asset classes thus

[20] Another reason why the issue is relevant for both income from DB and DC plans lies in potential cash-outs of pension rights upon job separation. As much as 60 percent of accumulated pension plans are currently cashed out upon job change (Yakoboski 1997). The cash-out rate is lower among accounts with high balances, so that only 21 percent of account dollars were cashed out. This remains a very sizeable amount of money with potentially large effects on both pension and non-pension wealth accumulation. Both DB and DC plans may be cashed out. An increasing fraction of DB plans (64 percent in 1993) provide the option of a lump-sum distribution (LSD) upon job separation. Among DC plans, 87 percent provides an LSD option upon job separation. Overall, 72 percent of plan participants were able to take an LSD in 1993, up sharply from 48 percent in 1983 (Scott and Shoven, 1996). Logical consistency with Task I-2 (income from assets and savings) requires that pension cash-outs are accounted for in Task I-3.

corresponds reasonably closely to the rates documented in Ibbotson Associates (1998) and O'Shaughnessy (1998). The literature offers some evidence on 401(k) portfolio allocation. See, for example, Yakoboski and VanDerhei (1996); Papke (1998); and Sundén and Surette (1998). *We recommend that stochastic variation across individuals and time be added to the rate of return on DC plan assets*, and that its distribution may be different from the distribution of rate of return on non-pension assets.

Note that variation in the rate of return may have far-reaching implications for income from private savings accounts which may be created if the Social Security system is (partly) privatized. Depending on the degree of freedom covered workers will have in directing their plan asset investments, potentially large differences in retirement income may result. The analysis of rates of return on DC plans will benefit the choice of distribution for the rate of return on private savings accounts.

Part I, Task 4: Develop predictions of partial retirement earnings

An important stochastic element in the prediction of partial retirement earnings is (partial) labor force participation. The Urban Institute letter report by Caroline Ratcliffe and Lawrence Thompson (29 May 1998) and Ratcliffe's presentation during the meeting with the Panel of Experts indicate that Urban Institute's preferred option includes a separate equation for labor force participation. Stochastic variation in this behavioral variable is introduced by drawing a random number.[21] We agree that this procedure generates sufficient variation and do not see other substantively important sources of stochastic variation.

Part I, Task 5: Project Social Security lifetime earnings

We did not identify stochastic elements that will have a substantial effect on variation in the distribution of Social Security lifetime earnings.

Part I, Task 7: Aging of retirement income and assets

Total income from all sources is the sum of individual components. All issues mentioned above apply, but we do not believe that any additional issues arise.

[21] Since labor force participation is part of the model, its variation falls into the "residual variation" category defined above. In Section 3.2.4 we suggest approaches for incorporating such residual variation.

Part II, Task 1: Demographic projections

Our models of marriage formation, marriage dissolution, and mortality do not impose any parameters from outside the model that may be subject to individual variation. We do not believe that any stochastic elements play a substantial role in determining the distribution of demographic outcomes.

3.2.4. Residual Variation

In making predictions, one tends to estimate the future mean and use it as the prediction. This implies that individuals with identical future characteristics (even if known with certainty) are assigned the same prediction, so that part of the variation in initial values is lost. Although not mentioned in the RFTOP, adding residual variation will be important to preserve the distribution of the projected outcomes. We now turn to the incorporation of residual variation. The discussion centers on types of models; each task or subtask may use one or more of these types of models.

Linear models with known residual distribution

The first approach applies to linear models in which the distribution of the residual is known. One then knows the theoretical distribution of the outcomes conditional on the forecast. A simple, but useful, example is the forecast of income based on a log-linear regression, in which the theoretical distribution resulting from residual variation is normal.

Consider the following simple log-linear model:

$$y_i = b'x_i + u_i$$

where y_i is log-income and u_i is distributed normally. Assuming that all future values of the covariates, denoted x_f, are known with certainty, the consistent projection of log-income is $y_f = \hat{b}'x_f$ and the consistent projection of income (accounting for the variance correction) is $\exp\left\{\hat{b}'x_f + \tfrac{1}{2}\hat{s}_u^2\right\}$. This projection, however, does not incorporate residual variation. Instead, draw a random number u_f from the normal distribution with mean zero and variance $\hat{s}_u^2$ and project future income as $\exp\left\{\hat{b}'x_f + u_f\right\}$. The resulting projection for individual *i* preserves the distribution in the projected population.

Whether the normality assumption holds for any of the income component models which the Urban Institute will estimate is an empirical question.

Linear models with measurable heteroskedasticity structure

A generalization of the known residual distribution discussed above is a residual distribution with known heteroskedasticity structure. For example, the distribution of wealth holdings is known to widen with age. If the residual structure of such outcomes is partly specified in terms of observables, the projection is a straightforward extension of the procedure described above. For example, suppose the residual in the model above is

$$u_i = h(z_i)v_i,$$

where $h(z_i)$ is a known function of observed covariates (such as age) and v_i is distributed *iid* normally, the projected outcome including residual variation is $\exp\left\{\hat{b}'x_f + h(z_f)v_f\right\}$.

Linear models with unknown residual distribution

While models are often specified and estimated under the assumption of a certain residual distribution, the actual distribution is typically unknown. In such cases, consistent projections may be obtained through the "smearing estimator," which explicitly accounts for potential misspecification of the distribution of the residual (Duan 1997). The method takes draws from the empirical distribution of residuals used in the estimation. In the log-linear example, the mean of $\hat{y}_f = \exp\left\{\hat{b}'x_f + u_f\right\}$ is no longer $\exp\left\{\hat{b}'x_f + \frac{1}{2}\hat{s}_u^2\right\}$. The smearing estimate of the mean is $\hat{y}_f = \frac{1}{N}\sum_{d=1}^{N}\exp\left\{\hat{b}'x_f + u_d\right\}$, where N is a (large) number of draws d from the empirical distribution of the residuals. Multiple random draws from the empirical distribution of residuals may be required to consistently estimate the mean forecast value.

This consistently projected mean, however, does not incorporate residual variation. The simplest method to account for residual variation is to only draw once from the empirical distribution of the residuals, so that the projected outcome is $\hat{y}_f = \exp\left\{\hat{b}'x_f + u_d\right\}$. The resulting projections preserve the original distribution of the outcomes, provided that the projection sample is sufficiently large. The 1931-1960 birth cohorts in the combined 1990 and 1991 SIPP samples should yield such a sufficiently large projection sample.

Tobit models

The residual distribution of Tobit models, including the fixed effect Tobit model proposed for modeling Social Security earnings, is by assumption normal. The projection of future values of the outcome (whether that be earnings or the ratio of earnings to the U.S.-wide average Social Security wage) is given by a straightforward generalization of the procedure outlined for linear models with known residual distribution, above:

$$\max\left[0, \min\left[t_f, \exp\left\{\hat{b}' x_f + u_f\right\}\right]\right],$$

where t_f is the future value of the maximum taxable income (or its ratio to the future U.S.-wide average wage). In other words, the projected income value is truncated from below at zero and from above at the maximum taxable wage, so that the correct fraction of individuals is projected to have zero or maximum earnings.

Probit and ordered probit models

Consider an (ordered) probit model for (partial) labor force participation after retirement, i.e., after the individual starts claiming OASI benefits. The probit index function is given by

$$p_i^* = a' x_i + w_i,$$

where $w_i \sim N(0,1)$ and participation is determined by the value of p_i^* relative to one or more thresholds. In a qualitative model like this, a consistent forecast is a participation rate that may not be encountered in the underlying data. Residual variation may be incorporated by drawing a future value of w_i from the standard normal distribution, say, w_f, and comparing $\hat{a}' x_f + w_f$ to the (ordered) probit threshold(s). The resulting projection will be an actual value in the underlying data.

We believe that the procedure proposed by Caroline Ratcliffe and Lawrence Thompson for labor force participation after retirement (as outlined in their Letter Report of May 29, 1998, and in the presentation on June 3, 1998) follows this approach.

Continuous-time hazard models

In continuous-time hazard models, the analogy of a consistent "mean" projection is the expected failure time, i.e., the expected duration of the spell. The expected failure time is given by

$$\hat{T} = \int_{t=0}^{\infty} t f(t) dt,$$

where $f(t)$ is the probability density distribution of the spell's duration. This density is by definition equal to the product of the hazard and survivor functions, $f(t) = h(t)S(t)$. As in all cases discussed above, the projection of expected durations (until marriage, divorce, widowhood, and death) does not preserve the distribution of such outcomes in the population.

One approach to incorporate residual variation is to compute transition probabilities over short duration intervals, draw a random number that is uniformly distributed over the unit interval, and project a transition on the basis of a comparison between the transition probability and the random number. For example, suppose that the probability of getting married for a person with certain characteristics in the next month is 0.02. We draw a random number between zero and one; if that number is less than 0.02, we project that the person got married in the next month. If not, we compute the transition probability in the following month, draw a new random number, et cetera. This is the procedure we discussed in our proposal.

An alternative approach is to draw a random number k between zero and one and compute the failure time $\hat{T}_f$ as the duration at which the survivor function equals the random number: $S\left(\hat{T}_f\right) = k$. The resulting duration follows the same distribution as the underlying sample distribution. This procedure requires formulation of the inverse of the survivor function; for piecewise-linear log-hazard functions (generalized Gompertz) as proposed for demographic transition models, this inverse has a closed-form solution. We anticipate that this second approach is computationally more efficient than the first. Another, minor, advantage is that it generates exact durations rather than time intervals.

3.3. Develop Approaches to Adding Economic Variability

3.3.1. Objectives

The overall objective of Task 2 is to ensure that the distributions of the income and demographic outcome variables are preserved in the projections. The goal of this subtask is to develop approaches for adding economic variability to the projections at the individual level for economic and fiscal variables that were identified in Task 2-1.

3.3.2. Overview

Uncertainty about future values of variables of interest stems from several sources. The first is *residual variation*: no statistical or economic model, no matter how detailed, is capable of fitting historical data perfectly.[22] There is always an implicit or explicit residual term; that term also applies to predicted values. A second source is uncertainty about the future values of the explanatory variables used to make the forecast. Those variables may be (time-varying) covariates or they may be *stochastic elements* such as the rate of return on assets, employer 401(k) match rates, etc. John Rust, in his letter of June 26, 1998, labels these *predetermined variables.* A third source stems from uncertainty about model parameters. Model parameters need to be estimated, are therefore uncertain, and widen confidence intervals of the forecasts. Rust speaks of *estimation noise*.

The objective of this task is to ensure that the distributions of projected outcomes correspond closely to the distributions of observed outcomes in the SIPP. The distributions need not and probably should not be identical: we would expect wages to grow over time, life expectancy to increase, etc. The RFTOP states that "the diversity of the American population ... should be maintained as the individuals are aged in the projection model." In our interpretation, the objective is thus mostly aimed at preserving the variance of projections.

Given this interpretation, we think that uncertainty about model parameters should not be considered. For example, an overestimate of the growth rate of wages results in higher projected wages for all simulants. It affects the mean projection, with only second-order effects, if any, on the variance. In other words, estimation noise is of limited relevance to diversity and inequality analyses. Indeed, as noted by Rust, "the amount of estimation noise is a third order problem compared to the first order problem of describing the uncertainty in the path of the predetermined variables." This is not to say that it would not be useful to simulate the model many times with different draws from the parameter distribution. Such an exercise would be a form of sensitivity analysis to the values of imprecisely estimated parameters.

[22] Except in a specific sample when the number of parameters is increased until zero degrees of freedom. Drawing a new sample would no longer result in a perfect fit.

To preserve the diversity of the population in the projections, we do need to account for residual variation and stochastic elements. The distinction is often a matter of model specification. If a variable explicitly enters a model, its variation causes stochastic variation. If it is omitted, its variation becomes part of residual variation. The distinction may have caused some misunderstanding. For example, our Letter Report on Task 2-1 (replicated here as Section 3.2) stated that we "do not believe that any stochastic elements play a substantial role in determining the distribution of demographic outcomes." Rust took issue with that statement, arguing that "there is considerable uncertainty about things like divorce, remarriage, death of a spouse, and so forth..." We fully agree and have no illusions that our model projects perfectly. What we intended is that all uncertainty is of residual nature; the explanatory covariates in our demographic models are all observed in the data and do not have uncertain future values (e.g., for education, race, and the like there is no source of stochastic variation). In other cases, (e.g., for future marital status transitions) these covariates are explicitly modeled and accounted for.

3.3.3. Residual Variation

As indicated in our Letter Report on Task 2-1 (Section 3.2), we propose that the Urban Institute and RAND both incorporate residual variation in their projections. Doing so is a very natural element of the projection procedure. Furthermore, if RAND were to add residual variation to Urban's projections, Urban staff will need to spend considerable time documenting every estimation and projection program, all data will need to be transferred, and all projection programs will need to be re-run. In other words, the fixed costs would far exceed the very low costs of adding residual variation. It would also delay the moment at which the Near Term Model may be applied to policy simulations, because projections without residual variation have more limited value for studying distributional consequences and inequality.

Section 3.2.4 specified how residual variation may be added. Note that no Monte Carlo simulations are involved; only a single random number needs to be added to every projected outcome. For binary choice models, the projection technique naturally and almost inevitably incorporates residual variation. For example, the decision to engage in part-time work after retirement is projected by adding a normally distributed random number to the probit index function to project whether an individual retiree will work. For linear and Tobit models, adding a random draw from the residual distribution is trivial. For hazard models, the issue is more complicated, though not more complicated than the computation of an expected duration. RAND will incorporate residual variation in its demographic hazard models.

Naturally, we will provide any assistance necessary to clarify the procedure.

3.3.4. Stochastic Variation

We propose three options for incorporating stochastic variation. All three assume that stochastic variables and their distributions have been identified, as discussed in Section 3.2. The first option is to replace the constant value of a stochastic element by a draw for every individual (and every time period or every employer or every marriage, as appropriate); the second is to conduct Monte Carlo simulations; and the third is based on a single large replication.

The RFTOP also mentions the Ibbotson Associates technique of estimating final values of variables such as savings and assets (Ibbotson Associates, 1998, Chapter 9). Our interpretation of that technique is that Ibbotson draws from a known statistical distribution (the book mentions the log-normal distribution as appropriate for asset return relatives) rather than from an empirical distribution. As such, we think that this establishes the distribution from which stochastic elements are to be drawn. It therefore relates to a preliminary step of incorporating stochastic variation. We will apply either technique as appropriate for the stochastic element under consideration.

Option 1: A Single Set of Draws

Since none of the income projection methods has been finalized, it is somewhat difficult to illustrate techniques through examples. We will make some assumptions in the following discussion; some of those may not apply.

In the first option, we replace a constant variable—which should really be subject to variability—by a single draw or set of draws. For example, the model that projects income from assets contains a parameter representing the rate of return on assets. The baseline model may set that parameter to a constant value for all individuals and all time periods. In the first option, we draw rates of return for every individual and every year from the distribution of annual rates of return (as specified in Section 3.2) and substitute it for the constant model parameter. The resulting distribution of incomes from assets will have a variance that is greater than the variance under the assumption of a constant rate of return (provided that the distribution of rates of return is not in some very systematic way negatively correlated with asset holdings).

If the simulation sample were infinitely large, this technique would result in the appropriate distributions of the outcomes of interest. The combined 1990 and 1991 SIPP simulation sample has roughly 50,000 individuals, which may not be sufficiently large to support appropriate distributions on certain rare subpopulations. The next options remedy this.

Option 2: Monte Carlo Simulations

A natural way to implement stochastic variation is through Monte Carlo simulations. Consider again projections of income from assets that may be strongly affected by the rate of return on assets. The projection program would be run many times, say, 100 times. The rate of return on assets would not be constant, but vary across individuals and across time periods. This is achieved by drawing a rate of return for every individual and every time period. As the projection model is run 100 times, 100 sets of rates of returns will be drawn, resulting in 100 different projections for every individual.

The results should be aggregated first over individuals and second over Monte Carlo iterations. Suppose one is interested in income inequality as measured by a Gini coefficient. Each Monte Carlo iteration results in income projections for all 50,000 individuals in the sample. These should first be aggregated into a Gini coefficient, resulting in 100 Gini coefficients. In the second step, the mean Gini coefficient and its standard deviation may be computed.

Option 3: One Large Replication

A third alternative essentially combines the first two options. One may take the simulation sample of 50,000 individuals and duplicate them a large number of times, say, 100 times. This results in a simulation sample of 5,000,000 individuals. Assign all individuals (sets of) draws from the distribution of the stochastic element and run the projection program. The resulting projections support analyses of distributional consequences even for rare subpopulations, because each member of such population is represented 100 times.

One virtue of this approach is its simplicity. Duplicating observations is trivial, and the projection program needs to be run only once. A disadvantage is that it does not yield an estimate of the variance of the summary statistic of interest. The Gini coefficient that is based on 5,000,000 individuals is equal to the mean coefficient obtained from Monte Carlo simulations (up to nonlinearities), but no standard deviation is obtained, i.e., the technique is not so rich in the information that it generates. Another disadvantage is that the technique may require substantial amounts of disk storage space. A minor advantage, finally, is that the resulting projection data support the computation of alternative summary statistics. A Monte Carlo simulation would need to be re-run, since the summary statistic would need to be computed for every iteration.

Table 3.2 summarizes the advantages and disadvantages associated with the three options. Two asterisks imply a higher level of attractiveness.

Table 3.2. A Comparison of the Three Options

	Ability to analyze rare subpopulations	Information richness	Simplicity	CPU time	Disk storage	Ability to generate alternative summary measures
1	*	*	**	**	**	**
2	**	**	*	*	**	*
3	**	*	**	*	*	**

We recommend using option 1 except in applications where rare subpopulations are being studied or where there is some other compelling reason to gain the efficiency of multiple replications. The organizational and coordination costs of performing multiple replications for Monte Carlo or large replication may usually out-weigh the benefits. If multiple replications are performed, retaining the replication number either option 2 or option 3 can be performed ex post.

3.4. Implement Technique for Adding Economic Variability

3.4.1. Objectives

The overall objective of Task 2 is to ensure that the distributions of the income and demographic outcome variables are preserved in the projections. Task 2-1 identified economic and fiscal variables that should be subject to stochasticity. Task 2-2 developed techniques for implementing such stochasticity into the projection models. The goal of Task 2-3 is to implement the technique as selected by the Task Manager for adding economic variability to the projections at the individual level.

3.4.2. Stochastic Variables

Section 3.2 identified the following economic variables as important for preserving the distribution of projected income flows and demographic states.

- The rate of return on assets and savings, for projecting retirement income from assets and savings (Part I, Task 2). We recommended that the projections be subject to rates of return that vary both across individuals and over time.
- The employer contribution match rate, for projecting retirement income from defined contribution pensions (Part I, Task 3). We recommended that the projections be subject to employer contribution match rates that are an inverse function of the employee contribution rate, and vary across individuals.
- The rate of return on DC plan account balances, for projecting retirement income from defined contribution pensions (Part I, Task 3). We recommended that the projections be subject to rates of return that vary both across individuals and over time.

At the time we identified these stochastic variables, no decision had been made on the structure of The Urban Institute’ s model underlying retirement income from assets and savings (Part I, Task 2). The SSA Task Manager opted for a model in which age-wealth profiles are estimated and projected and in which the rate of return on assets and savings no longer explicitly enters. Implicitly, of course, the rate of return still plays a central role, but year-on-year and cross-individual variation in the rate of return is now part of the residual structure.

Our reports emphasized that, in addition to stochasticity of economic and fiscal variables, the residual structures of the various component models in MINT play an important role in maintaining the distribution of income and demographic outcome variables of future projections. We further discuss residual variation in the next section.

3.4.3. Techniques

Sections 3.2 and 3.3 distinguished residual variation from stochastic variation. Both play an important role in maintaining distributions of outcomes in projections. For practical purposes, the distinction is largely a function of model structure and specification. If a certain parameter enters the model and is assumed fixed (but really varies across individuals, over time, or otherwise), projections will exhibit too little variation. If a parameter is omitted from the model, its variation is absorbed by the residual term; projections that set future residual terms to their mean value (usually zero) will therefore again exhibit too little variation.

We recommended that both The Urban Institute and RAND include residual variation into their projections. This issue was further discussed in a meeting between Howard Iams, Lee Cohen, Eric Toder, Gary Burtless, and Stan Panis on 20 July 1998. Iams subsequently directed The Urban Institute and RAND to add single draws from residual distributions to each party' s projections. Details on appropriate procedures for various types of models are provided in Section 3.2.

In Task 2-2 we developed three techniques for incorporating stochastic variation: (1) substitute a single draw from the stochastic variable' s distribution for its mean value; (2) repeatedly substitute draws from the stochastic variable' s distribution for its mean value, and compute aggregates of the projected outcomes of interest (known as Monte Carlo simulations); and (3) replicate the simulation sample and draw once from the stochastic variable' s distribution, essentially leading to similar results as the second option. We recommended the first option: draw once from the stochastic variable' s distribution and use it, instead of a constant value, to project the outcome of interest.

As stated above, projections of income from assets and savings no longer involve a rate of return. The two remaining stochastic elements that we identified as important, the employer match rate of DC plan contributions and the rate of return on DC plan balances, both apply to The Urban Institute' s Task 3 (retirement income from pensions). The SSA Task Manager directed The Urban Institute to draw single values from the empirical distributions of employer DC plan contribution match rates and rates of returns on stocks and bonds and to substitute these draws into the projection model.[23] This directive is consistent with our recommendation.

[23] The distribution of employer match rates is based on 1995 Survey of Consumer Finances (SCF) data; the distribution of rates of return on the stock portion of DC plan balances is based on S&P 500 returns from 1952-1994; the distribution of rates of return on the fixed-income portion of DC plan balances is based on T-Bill rates over the same period. The rates of return are assumed to be normally distributed (in slight deviation from Ibbotson, 1998, but consistent with Cohen, 1998) with means equal to variances. At the time of this report, the exact parameters have not yet been decided upon. Cohen (1998) found 1952-1994 average real rates of return for the S&P 500 stock index of 6.98 percent and 1.17 for T-Bills. See Letter Report for Task 3-2 of The Urban Institute.

The Optimal Number of Draws

Christopher Bone, in his memorandum to Howard Iams of July 24, 1998, raised an important and intellectually challenging issue. Section 3.3 asserted that a single set of draws to replace constant variables is appropriate except where rare subpopulations are being studied or there is some other compelling reason to gain the efficiency of multiple replications. Bone pointed out that this does not provide those using the model with any guidance on what constitutes a "rare" subpopulation. Furthermore, cumulative mortality increases the number of relatively rare subpopulations: a subpopulation may not be "rare" at age 62, but may be "rare" in 2020.

Note that Bone's critique applies to residual variation as well as to stochastic variation. The Urban Institute and RAND both include a single draw from the residual distribution into the projections. Ideally, to get the full distribution, very many draws should be included.[24]

Bone went on to request the development of a method for classifying the number of replications needed to ensure sufficient variation in the population.

The issue is important not only for its relevance to stochastic variation, but also because it generalizes to MINT as a whole: "*What is the smallest subpopulation for which MINT is capable of generating reliable distributional consequences?*" Obviously, there is no hard and uniformly applicable answer to this question. A core feature of any statistical prediction model is that its predictions have smaller confidence intervals and larger power the larger the subpopulation of interest is. Rather than solve the issue in general terms, we now discuss the factors relevant to the subpopulation size issue.

For discussion purposes, consider the following thought experiment. A MINT user wants to figure out whether a certain policy measure (or combination of policy measures) results in an increased poverty rate by 2020 among, say, elderly African-American women living alone. Assume for now that we follow the best possible procedure (within more fundamental limits of the MINT component models), namely to draw very many times for both stochastic variables and residuals. The user runs MINT twice: once under baseline assumptions (no change in policy) and once incorporating the package of policy measures. Suppose, for argument's sake, that the baseline poverty rate in 2020 among the subpopulation of interest is projected to be 20 percent ($\hat{p}_1 = 0.20$) and the post-policy rate 23 percent ($\hat{p}_2 = 0.23$). These rates have distributions, because the parameters underlying MINT are not known with certainty, but have been estimated. We are interested in the change in poverty rate ($d = p_2 - p_1$), for which we have a point estimate of 3 percent ($\hat{d} = 0.03$). Its standard

[24] The issue is actually less relevant for stochastic variation of rates of return than for other variation. For projecting the value of a DC plan upon retirement, the relevant rate of return is the rate as it compounds over all accumulation years. Since a rate is drawn for every year anew, the compounded rate is very stable.

deviation may, in principle, be estimated.[25] The standard deviation depends inversely on the size of the subpopulation.

To determine required subpopulation sample sizes, two types of error are relevant.

The first type of error, known in the statistical literature as "Type I error," arises when the null hypothesis is true but rejected by the simulation. Suppose the policy measure under simulation does not affect the poverty rate of the subpopulation of interest, i.e., the null hypothesis that $d = 0$ is true. Whether the point estimate, $\hat{d} = 0.03$, is large enough to reject the null hypothesis depends on the standard deviation of $\hat{d}$ and the chosen significance level. In practice, a probability of a Type I error of 5 percent tends to be considered acceptable. That 5 percent probability ("significance level") corresponds to 1.96 standard deviations, so if 0.03 exceeds 1.96 times the standard deviation of $\hat{d}$, the MINT user would conclude that there is a significant increase in the poverty rate, and thus make a Type I error.

The second type of error, known as "Type II error," arises when the alternative hypothesis is true but rejected. This type of error is directly related to the "power" of a test. Power is defined as one minus the probability of a Type II error. To assess the power of a test, the MINT user must take a stance on an alternative hypothesis of interest. For example, the user may decide that increases in the poverty rate of less than 4 percentage points are of no great concern, but that the model must be capable of detecting changes of 4 percentage points or more. This corresponds to an alternative hypothesis that $d = 0.04$. Would the user reject that hypothesis? The answer depends again on the standard deviation of $\hat{d}$ and the probability of a Type II error that the user is willing to accept. In practice, a probability of a Type II error of 20 percent tends to be considered acceptable; put differently, users tend to require a power of at least 80 percent. This 20 percent probability corresponds to 0.84 standard deviations, so if 0.04-0.03=0.01 exceeds 0.84 times the standard deviation of $\hat{d}$, the MINT user would reject the hypothesis that the poverty rate increased by 4 percentage points, and thus make a Type II error.

Note that both types of errors depend on the size of the subpopulation of interest. Larger sample sizes reduce the probabilities of making Type I and Type II errors. In

---

[25] Note that d is a function of model parameters, q, characteristics of the subpopulation, X, and policy parameters. Policy parameters and respondent characteristics are assumed to be known with certainty, so we may approximate the variance of d by:

$$\hat{s}^2_d \approx \frac{\partial d}{\partial q'} \hat{\Sigma}_{qq} \frac{\partial d}{\partial q},$$

where $\partial d / \partial q$ is the (numerically computed) first derivative of the increase in poverty rate with respect to MINT parameters and $\hat{\Sigma}_{qq}$ is the covariance matrix of the model parameter estimates. The latter, $\hat{\Sigma}_{qq}$, is not typically reported in model estimates, but is accessible in most software packages. It is block-diagonal in the MINT model, because model modules have been estimated separately, under the assumption of independence. The computation will be hugely difficult and time-consuming, but is possible.

principle, it is possible to compute the sample sizes that are required to achieve a significance level and power that are selected as sufficiently high.

The above assumed that we conducted Monte Carlo simulations with infinitely many draws for both residual and stochastic variation. As the number of draws goes down to a finite number, and perhaps as low as one (such as is currently the case), the standard deviation of $\hat{d}$ becomes subject to uncertainty about residual and stochastic values, in addition to the uncertainty from imprecisely estimated parameters. Its computation becomes extraordinarily complex.

Approaching the issue from the opposite direction, we investigated the number of respondents in the simulation sample that represent subpopulations of actual policy interest. The combined 1990, 1991, 1992, and 1993 SIPP panels contain 65,369 full panel respondents born in 1931-1960. Of these, 17,998 are projected to become deceased before the year 2020, and 47,371 (72.5 percent) are projected to survive. Table 3.3 shows the number of respondents among the 47,371 survivors that are part of subpopulations of interest. To indicate the political clout of these subpopulations, the table also shows population-weighted figures, i.e., the numbers of Americans that are represented by the SIPP respondents.[26]

Table 3.3. Selected subpopulations in the year 2020

	Simulation sample size	Weighted to U.S. population
African American women living alone age 65-89	1,858	2,912,285
Native American women living alone age 65-89	108	164,778
Hispanic women living alone age 65-89	1,172	1,594,340
African American women living alone age 75-89	747	1,095,386
Native American women living alone age 75-89	48	76,014
Hispanic women living alone age 75-89	504	632,902
Divorced women who have been married 10+ years	7,962	11,920,746
Individuals age 85-89	2,213	3,174,837

Note that the numbers of SIPP respondents that are members of these subpopulations and survive through the year 2020 are fairly high. For most subpopulations we have more than 500 sample members. At issue is the number that is required to determine a distribution, or, of keener political interest, the fraction of individuals below the poverty line. Without working through the formal model outlined above, it is impossible to know such minimum sample sizes. Intuitively, one may feel comfortable that 50-100 observations are sufficient to pin down changes in a poverty rate, especially if poverty is a relatively common occurrence, such as it is among elderly Native American women.

[26] Note that the oldest person, born in 1931, is only 89 years old as of January 1, 2020. As much as 32 percent of the individuals born in 1931-60 are expected to live past their 90-th birthdays, so projections of the oldest-old may gain precision from incorporating pre-1931 cohorts in the projections.

4. Consistency Checks

4.1. Interdependencies Between Components of Retirement Income

4.1.1. Objectives

Task 3 is concerned with both internal and external model consistency. Internal consistency refers to corrections for correlation among income components that are projected separately. External consistency refers to consistency of summary statistics[27] from model projections with external macroeconomic or other models that project similar summary statistics. This task applies to both our own projections and those of The Urban Institute and Brookings Institution.

Subtask 3-1 is to identify the components of each individual' s retirement income that are modeled separately, assess whether they should be consistent with each other, and recommend techniques to make them consistent.

4.1.2. Overview

A number of potential behavioral and/or correlational interactions exist between the various components of retirement income. By far the most important issue for any policy simulation exercise is that potential behavioral responses to change may not be reflected in projections of retirement income taken one at a time and assuming exogeneity of other sources of income, for example. The modeling and estimation of the relevant behavioral relationships are the subject of current debate and should be the subject of further research. They are outside the scope of the current effort but should be kept in mind and improved whenever possible. We focus here on a discussion of more correlational interactions among sources of retirement income, but even these are not fully understood and may be outside the scope of the current effort. Our purpose is to raise the issues for discussion and potential resolution or postponement for later analysis.

4.1.3. Discussion

A number of sources of positive correlation exist between various sources of retirement income, which, if ignored and if each source of retirement income is treated as independent, will potentially bias analyses of the distribution of retirement income, incidence of poverty, and like calculations.

[27] There is one more important consistency, namely between the SIPP (and/or PSID) and the U.S. population. While both surveys aim to be representative of the population, discrepancies between survey aggregates and U.S. aggregates may arise from the way questions are asked and other sources. It is the responsibility of The Urban Institute/Brookings Institution and ourselves to adjust parameter estimates for such discrepancies before applying them in the projections. We anchor the demographic transition models to vital statistics, as described in detail in Panis and Lillard (1996).

Retirement income is strongly affected by, if not largely determined by, pre-retirement behaviors and outcomes, which may induce correlation in sources of retirement incomes. And post-retirement behavior, such as saving or the lack thereof, may reflect similar pre-retirement behavioral patterns before.

Dimensions of pre-retirement behavior and outcomes that affect retirement income include: (1) life cycle patterns of work and earnings; (2) marriage, divorce, and the work and earnings of spouses; (3) work-related fringe benefits, including health insurance and DC and/or DB pension accumulation; (4) Social Security contributions and benefits; and (5) pre-retirement saving and financial asset/wealth accumulation.

Because these pre-retirement outcomes are obviously important determinants of financial resources — assets and income — in retirement, important interactions among pre-retirement outcomes will affect the distribution of retirement income. While the following discussion raises a number of potentially important interactions from the literature, the current state of knowledge and agreement on these issues does not permit easy incorporation into the MINT modeling. Instead, it suggests important issues for further research, and potentially feasible adjustments to the current effort may be forthcoming from discussion of the issues as work progresses.

There are a number of relevant relationships from the literature (published and not yet published). An obvious first relationship is that both Social Security and pension benefits are related to the level and pattern of pre-retirement life cycle earnings. Lillard and Weiss (1997), for example, report a positive relationship between post-retirement income from Social Security and from pensions using the old Longitudinal Retirement History Survey data. Hurd, Lillard and Panis (1998) show that workers with smaller retirement accounts are more likely to cash-out the account and spend the money when changing employers near retirement. In addition, some DB pensions are reduced if the beneficiary receives Social Security benefits. The SIPP does not offer this type of information; in the Health and Retirement Study (HRS), about 15 percent of respondents indicated that their future private pension would be offset by Social Security benefit receipts. The opposite may also apply: A Social Security benefit payable to a (divorced/surviving) spouse may be reduced if the person receives a periodic payment based on his or her own employment that was not covered under Social Security from the Federal Government, a State, or a political subdivision of a State (§ 1836, Social Security Handbook 1997).

Marital status is clearly important, and not only because of the spouse and survivor benefits of Social Security. Smith (1997) reports greater assets and asset accumulation in the PSID for persons married over a five-year period than for unmarried persons over the same period. Unpublished results by Lillard and Karoly (1997) show that for men, household wealth over the life cycle is strongly positively related to their own permanent earnings and only weakly related to the permanent earnings of the women to whom they are married. On the other hand, for women, household wealth over the life cycle is strongly positively related to the permanent earnings of the men to whom they are married and only weakly related to their own earnings. As a result, retirement assets and income from assets will be positively related to marital history.

The level of saving and the form of saving are clearly important. Pre-retirement and post-retirement consumption and saving behavior are likely to closely related, so that those persons who do not save for retirement are least likely to maintain assets after retirement, and there are substantial proportions of the population with little or no assets at retirement. Similarly, those persons with little or no savings at retirement are likely to have low wage rates, so continuing to work is a relatively unattractive option and their Social Security replacement rates are high, as noted by Hubbard, Skinner and Zeldes (1997). The form of saving may also be important. For example, 401(k) assets may be offset by mortgage debt (Engen and Gale 1997). However, there is some controversy around that issue. See, for example, Poterba, Venti, and Wise (1995); Venti and Wise (1996); and Poterba, Venti, and Wise (1996).

There also may be interactions between work and Social Security or pension benefits. Social Security benefits are reduced if the beneficiary's earnings exceed a certain threshold. For example, beneficiaries age 65-69 may earn $14,500 in 1998; for each $3 in additional earnings, their benefits are reduced by $1. While Social Security benefits may be reduced based on the beneficiary's earnings, they may be taxed based on any income. Social Security benefits are partially subject to income taxation for higher-income retirees. Beneficiaries with incomes of more than the base amount ($25,000 if single and $32,000 if married) are liable for income taxes on a portion of their OASI benefits. This correlation applies to after-tax income (Task 4), but also has behavioral implications, as in the next item. Partial taxation of Social Security benefits on the basis of other income components (e.g., from assets) may reduce incentives for labor force participation and thus earnings among the elderly (Part I, Task 4). Conversely, a reduction in Social Security benefits (which may be simulated as part of a proposed policy change) may induce higher labor force participation and earnings among the elderly.

4.2. Identify Macroeconomic Models and Their Use for Benchmarking

4.2.1. Introduction

This subtask identifies appropriate macroeconomic forecasting models for validating, and possibly benchmarking, the MINT microsimulation model.[28] The goal of this subtask is to support analyses of the extent to which aggregate forecasts implicit in MINT are consistent with accepted, external forecasts.

Linking micro- and macromodels at some level is fairly common.[29] For example, a non-central part of a microsimulation model may rely on simple time-series techniques to forecast a necessary macro-level variable. Alternatively, a macroeconomic model may be fully imbedded into a microsimulation model so that behavioral feedbacks between the models may exist. For this subtask, we focus on a third type of link that uses a macroeconomic forecasting model to validate the implicit aggregate forecasts of the microsimulation model.[30] If the forecasts are close by some metric, then the microsimulation model is considered "valid."

We proceed as follows. Section 4.2.2 describes the characteristics of the ideal macroeconomic forecasting model for validating MINT, without concern for whether such an ideal actually exists. We conduct a survey of macroeconomic forecasting models and select the most promising candidate models in Section 4.2.3. Section 4.2.4 discusses the mechanics of aggregating, validating, and benchmarking MINT.

4.2.2. Macroeconomic Model Characteristics

Before we examine potential macroeconomic models, we first list the characteristics of a macroeconomic model that would be appropriate for validating MINT. It should not be expected that a macroeconomic forecasting model will contain all these characteristics; rather, these criteria will serve as a guide in the selection of an appropriate macroeconomic forecasting model.

The ideal macroeconomic model for our purposes possesses the following characteristics. The model must:

- *Forecast through 2020*. MINT forecasts retirement income through the year 2020. A macroeconomic model that serves to validate MINT projections must therefore be specified to forecast through at least the year 2020.

---

[28] We gratefully acknowledge substantial expert input from Steven Haider.
[29] Anderson (1990) provides a detailed discussion of different types of linkages with numerous examples.
[30] Sargent (1985) provides a detailed discussion of general validating procedures for simulation models.

- *Have been evaluated for accuracy.* The purpose of the validation procedure is to compare MINT to macroeconomic forecasts that are considered "good." The macroeconomic model must therefore have a proven track record, for example, through comparisons of past projections with actual outcomes.
- *Forecast different income types.* Most macroeconomic models forecast aggregate quantities such as GDP. By contrast, MINT will project individual income components as well as aggregate retirement income. Ideally, the macroeconomic model should forecast the same types of personal income, such as income from earnings, pensions, Social Security, savings, etc.
- *Forecast by cohort.* MINT is specified to forecast retirement income of the 1931-1960 birth cohort only. The ideal macroeconomic model therefore forecasts income by birth cohort, permitting a direct comparison.
- *Provide confidence intervals.* A formal statistical test of whether two forecasts differ requires confidence intervals (i.e., standard errors) for both forecasts.
- *Be fully documented.* Projections of the macroeconomic model and MINT may only be expected to be identical if their underlying assumptions are the same. Where identical assumptions are lacking (as will very often be the case), discrepancies may arise. An evaluation of the source of such discrepancies requires full knowledge of both models' underlying assumptions.
- *Be in the public domain or owned by SSA.* Many macroeconomic forecasting models are maintained by private, for-profit consulting firms. These firms tend to charge considerable sums for the use of their models, which in itself is an argument for preferring public-domain models. More importantly, a model that is in the public domain or owned by SSA is more likely to permit a thorough examination of the underlying structure than proprietary models. It may furthermore be more readily adapted to serve MINT's needs.
- *Support projections under alternative policy regimes.* MINT's purpose is to permit scenario analysis of alternative policy reforms. The ideal macroeconomic model is capable of forecasting under these same alternative policies, so that both the baseline and the alternative projections may be validated. Preferably, no assistance from the model's keeper is required to project under alternative regimes.
- *Include behavioral feedbacks.* To ensure the feasibility of developing a microsimulation model in a timely fashion, the MINT architects have accepted a minimum number of behavioral feedbacks. The ideal macroeconomic model includes behavioral feedbacks, so that it supports future extensions of MINT and generates a rough indication of the bias introduced by the lack of behavioral feedbacks in MINT.

Although we do not think it is useful to explicitly rank the importance of each of these characteristics, we do consider it useful to indicate which characteristics are most critical. Specifically, an appropriate macroeconomic model for validating MINT should at least (a) forecast through 2020, (b) have been evaluated for accuracy, and (c) be fully documented.

4.2.3. Survey of Forecasting Models

Types of Macroeconomic Forecasting Models

Building on Brayton et al. (1997), we distinguish three types of forecasting models.[31] The first type of forecasting model is the traditional "IS-LM" (or Keynesian) macroeconomic model. These large-scale models specify the relationship between the macro-level variables, often explicitly including sluggish price adjustment. IS-LM models have been quite successful at forecasting the quarter-to-quarter performance of the economy. Examples of these types of models are the Fair Model (maintained by Ray C. Fair at Yale University) and the Washington University Macro Model (maintained by Macroeconomic Analysts).

The second type of forecasting model are "modern macro" models or "small scale macro models."[32] These models are characterized by the inclusion of optimizing agents with expectations that are explicit and rational. They tend to exclude certain sectors of the economy to remain tractable, and often feature only a limited number of policy levers. Although these models do not tend to forecast short-term fluctuations in the economy as well as IS-LM models, they are well equipped to forecast the long-term impact of policy regime changes.

The final type of forecasting model are largely statistical models. Such models fully capture the dynamic relationship among a few variables by relying on vector auto-regressive (VAR) techniques for estimation and projection. However, these models include few, if any, economic relationships. They tend to forecast the short-term quite well but are unable to forecast the effects of policy changes.

These three categories are not mutually exclusive. Many forecasting models will rely on model characteristics from different categories, depending on the particular forecasting goals. For example, the Federal Reserve Board's forecasting model for the United States (FRB/US) includes an underlying IS-LM model for the macro economy and an explicit characterization of expectations. Furthermore, many forecasting models will use a largely statistical model for aspects of the economy that are secondary to the modeling goals.

For purposes of MINT validation, the most promising models belong to the IS-LM class. They are most often used for forecasting, including by the Federal Reserve Board, the Congressional Budget Office, and the Council of Economic Advisers. They tend to perform quite well. However, the models tend to have at least one major drawback (in addition to often being proprietary): They usually include a "judgmental adjustment" or "fudge factor" that is to a large extent arbitrary and precludes the calculation of standard errors.

---

[31] See Fair (1994) for a historical review of macroeconomic forecasting models.

[32] See Leeper and Sims (1994) for a particular small-scale modern macro model. Also, real business cycle (RBC) models are considered small-scale modern macro models.

Potential Macroeconomic Forecasting Models

We surveyed the following macroeconomic models.

- BC: The Blue Chip Financial Forecasts are a consensus forecast of leading private economists and analysts. They forecast major financial characteristics of the United States such as GDP, Federal Funds Rate, and the Prime Rate. More information may be found on their web-site (http://www.bluechippubs.com).
- CBO: The Congressional Budget Office forecasts are used for federal governmental activities such as revenue and deficit planning. The CBO 10-year forecasts are readily available on the web (http://www.cbo.gov/reports.html). The CBO generates its forecasts both from its own analysis as well as using the forecasts from other models (DRI, MA, BlueChip and WEFA). See Congressional Budget Office (1998) for more details.
- COREMOD: This model uses the WUMM (see below) as its basis for short-run forecasts and uses a neoclassical growth model with representative households and firms (and myopic expectations) for long-run forecasts. The model was developed and is maintained by Macroeconomic Associates, LLC (http://macroadvisers.com); the forecasts are provided on a fee-for-service basis.
- DRI: This forecasting service is run by Standard and Poor (http://www.dri.mcgraw-hill.com). Five-year forecasts are publicly available.
- FAIR: The Fair model was developed by Ray C. Fair at Yale University. The model is a large-scale IS-LM model. Notably, the model is extensively documented, in the public domain, and available on-line for the public to generate its own forecasts (http://fairmodel.econ.yale.edu). See Fair (1994) for a detailed description of the model.
- FRB/US: The Federal Reserve Board model for the U.S. was developed at the Federal Reserve Board in the mid-1990s. It is a large-scale macro model that relies on an IS-LM framework for short run behavior, and it includes explicit expectations and optimizing agents for longer-run behavior. See Brayton et al. (1997) for a historical perspective on the development of the model and Brayton and Tinsley (eds., 1996) for a more technical description of the FRB/US model.
- MDM: The Macroeconomic-Demographic Model was developed as a microsimulation model usable for policy analysis by the National Institute on Aging in the 1980s. The MDM completely integrates a macroeconomic forecasting model developed by Dale Jorgenson and Edward Hudson. The model is capable of generating forecasts at least 75 years into the future and was used for extensive policy analysis during the 1980s. See National Institute on Aging (1984) for further details.
- RSQE: The RSQE model is maintained at the University of Michigan. It generates forecasts for 3 years. More information is available on its website (http://rsqe.econ.lsa.umich.edu).
- SSA: The Social Security Administration generates economic and demographic forecasts that it regularly uses to evaluate the long-run solvency of its various programs. For an example of its forecasts, as well as for information on how they are generated, see Board of Trustees (1997).

- WEFA: This group was formerly the Wharton Econometric Forecasting Associated established at the University of Pennsylvania by Lawrence Klein. The group merged with Chase Econometrics in 1987 to become WEFA; see its website (http://www.wefa.com). It provides forecasts for up to 25 years. The forecasts are generated with a traditional IS-LM, large-scale model.
- WUMM: The Washington University Macro Model was developed at Washington University and is now maintained by Macroeconomic Associates, LLC (http://macroadvisers.com). Its forecasts are provided on a fee-for-service basis. The model is a traditional IS-LM model that generates forecasts for 10 years.

Table 4.1 summarizes the basic characteristics of the major forecasting models that we examined.

Table 4.1. Forecasting Models

Model Name and Source	Acronym	Fully documented?[a]	Public domain?	Projection horizon
Blue Chip Consensus Forecasts	BC		No	5 years
Congressional Budget Office	CBO		Yes	10 years
COREMOD, Macroeconomic Advisors	CORE-MOD	Yes	No	>50 years
Standard and Poor's DRI	DRI		No[b]	5 years
Fair Model, Ray C. Fair at Yale University	FAIR	Yes	Yes	5 years
Federal Reserve Board, U.S. Quarterly Model	FRB/US	Yes	Yes[c]	>50 years
Macroeconomic-Demographic Model	MDM	Yes	Yes	>50 years
RSQE, University of Michigan	RSQE		No[b]	3 years
Social Security Administration Forecasts	SSA	Yes	Yes	80 years
Wharton Econometric Forecasting Associated	WEFA		No[b]	25 years
Washington University Macro Model, Macroeconomic Advisors	WUMM	Yes	No	10 years?

Notes: [a] We only mark "yes" for models for which we have acquired and examined detailed documentation. Such documentation may be available for other models, but because of undesirable model attributes (such as too short a forecast period), we have not acquired the documentation.
[b] Although some aspects of the forecasts are publicly available, the forecasts are compiled by a private company on a fee-for-service basis. Complete information will likely have to be purchased.
[c] Technically, this model is in the public domain. Realistically, assistance will be required from staff members at the Federal Reserve Board.

The table does not include a column with an indication of accuracy of historical model predictions. All models claim to be accurate, though few back it up in the documentation to which we have access. In addition, most evaluation period are very short (typically six quarters or less), even for models that forecast over long periods.

Recommendation

Obviously, none of the models satisfies all desired criteria as outlined in Section 4.2.2.

The COREMOD model appears reasonably well-suited for validation purposes. Among its disadvantages are that it (as many other IS-LM based models) includes a fudge factor to arrive at "desirable" predictions, and that it is not in the public domain.

The FAIR model does not include a fudge factor, is superbly documented, fully in the public domain, and executable on the Internet. Its main disadvantage is that it only projects out five years into the future. It is not always clear to us what criteria underlie the number of years that models chose to project out. With today's computing resources, there should not be any technical reasons. IS-LM models that provide long-run projections tend to use the IS-LM structure for a limited number of years and adopt broad trends thereafter. Perhaps the FAIR model may be adapted to support longer projection periods, but this would require assistance from Ray C. Fair, its architect.

The FRB/US model does not use a fudge factor to arrive at desirable predictions and it does provide standard errors of its predictions. Our conversations with Federal Reserve Board staff members indicated that the FRB is eager to see external application of its models, and willing to provide assistance.

The SSA projections are not based on an explicit model. SSA considers projections generated by other models and bases its own projections on the OASDI Trust Funds Board's best estimate of the future course of the population and the economy. It provides three projections termed "intermediate," "low cost," and "high cost." The intermediate projections represent the Trustees' consensus expectation of moderate economic growth through the projection period.

Based on the minimum requirements of forecast horizon through 2020, proven accuracy, and extensive documentation, we recommend the FRB/US and COREMOD models. In addition, we recommend comparisons with the SSA forecasts to ensure consistency across projections developed by SSA actuaries and MINT.

4.2.4. Validating and Benchmarking MINT

In this section, we address the mechanics of validating and, if necessary, benchmarking MINT. This section applies to the use of macroeconomic models and their use for benchmarking, so we will restrict the discussion to income flow projections. Section 4.3 treats validation of demographic projections using macro models that are not economic in nature.

We first describe issues in aggregating MINT's micro forecasts to the national level. Second, we draw attention to the MINT sample universe and contrast it to universes to which macroeconomic models apply. Third, we discuss the statistical and heuristic validation of MINT results vis-à-vis macroeconomic forecasts. Finally, we suggest adjustments that may be made to MINT projections to correct for discrepancies from macroeconomic forecasts.

Aggregation

In principle, aggregation of micro projections to the national level is achieved by simply computing a weighted sum:

$$Y^M = \sum_i w_i y_i \,,$$

where y_i is a projected vector of outcomes for individual *i* (such as of various income components or total retirement income), w_i is the (scalar) weight for individual *i*, and Y^M is the aggregated outcome vector. Aggregation of microsimulation projections to the national level thus requires proper weights, w_i.

MINT is based on the 1990, 1991, 1992, and 1993 SIPP panels. Projections are created only for individuals born in the years 1931-1960, and only for full panel respondents, i.e., only for those who responded to all waves of a particular panel. The Census Bureau provides weights for these simulants such that the weights for each panel add up to the covered sample universe.

The fact that MINT is based on multiple SIPP waves slightly complicates the weighting procedure, since untransformed weights would result in a weighted population of about four times the actual sample universe. Since each panel is designed to be representative of its sample universe, the normalization procedure makes no difference in expectation of Y^M, only in the efficiency of the projection (variance of the estimate Y^M). Simply dividing all weights by four to account for four panels yields an unbiased but inefficient estimate of aggregate income flows. The most efficient projection is obtained by weighting proportional to sample size in the projection. Specifically, let n^j denote the number of respondents in SIPP panel *j* (*j*=90,…,93). The most efficient transformed weights are:

$$w^*_{ji} = \frac{n^j}{n^{90} + n^{91} + n^{92} + n^{93}} w_{ji} \,,$$

where w_{ji} is the full panel weight for individual *i* in SIPP panel *j*. Table 4.2 shows sample sizes for the four SIPP panels and the corresponding most efficient weights. The 1991 SIPP panel was smaller than the others and its weights (`pnlwgt`) are largest, on average, so that the efficient weight factor is smallest. The table only includes individuals with positive full panel weight, `pnlwgt`, and only individuals born in 1931-1960. For sample selection details, see Section 2.6.

Table 4.2. Simulation Sample Sizes and Efficient Weight Factors

SIPP panel	sample size	weight factor
1990	16,821	0.2825
1991	11,914	0.2001
1992	15,491	0.2602
1993	15,311	0.2572
Total	59,537	1.0000

Comparability

MINT projections are based on the SIPP (with certain exclusions), which is representative of the U.S. civilian noninstitutionalized population as of the SIPP survey years. In accordance with the definition of Gross Domestic Product (total output produced within the borders of a country), macroeconomic models generally apply to all residents of United States territory. There are thus several discrepancies between the SIPP sample universe and the sample upon which most macroeconomic models are based. These discrepancies are likely to lead to discrepancies in projections.

- In 1990, approximately 600,000 individuals were in military quarters (1990 Census). These are excluded from the SIPP, which covers the civilian population only, but included in the universe underlying macroeconomic models.
- In 1990, approximately 3.3 million individuals were institutionalized. Of these, 1.75 million were in nursing homes, presumably not many from among the 1931-1960 birth cohorts. However, 1.13 million were incarcerated and another 0.44 million were in mental hospitals, juvenile institutions, or other institutions (1990 Census). The institutionalized population is excluded from the SIPP, but included in many macroeconomic models.
- MINT projections only account for individuals living in the United States as of the SIPP baseline survey. They thus exclude income from immigrants that enter the United States between the baseline survey years and 2020. During the 1980s, the annual gross inflow of legal immigrants was around 1.0 million, and the annual net inflow around 0.8 million (McCarthy and Vernez, 1997). Future flows are, of course, highly dependent on immigration policy. The Immigration Act of 1990 substantially increased the number of legal immigrants permitted starting in 1992.

Each of these discrepancies points at population counts and thus income flows that are smaller in MINT than in macroeconomic models. Over a 25-30-year period, the largest discrepancy probably stems from immigration. Its magnitude is difficult to ascertain, since assumptions on immigration policy in most macroeconomic models are not explicitly specified. An exception is SSA projections, which assume total annual net legal immigration to rise to 900,000 by the year 2000 and remain constant thereafter (intermediate scenario). COREMOD simply features an exogenous parameter for total annual population growth (without distinguishing birth and immigration).

Two additional comparability issues arise. First, ideally, we would like to separately validate MINT projections of income components (from partial labor force participation, from pensions, from assets, etc.), but to our knowledge, no macroeconomic model provides projections of income components. For practical purposes, validation should thus be restricted to MINT' s aggregate retirement income projections.

Second, since MINT only takes the 1931-1960 birth cohorts into account, aggregate MINT projections cannot be directly compared to macroeconomic model projections, which do not support projections by cohort. A first order solution inflates aggregate MINT projections for the 1931-60 cohorts to be reflective of the entire population. To avoid problems stemming from the changing age distribution, this is best done by computing the income position of the elderly relative to the general population as of the SIPP survey years and by multiplying the projections by the inverse of that relative income. A more sophisticated inflation procedure would account for temporal changes in the relative income position of the elderly.

Validation

Many statistical tests are available for testing the equality of two vectors of outcomes, such as one generated by MINT and the other by a macroeconomic model. Consider the following test statistic:

$$Q = \left(Y^M - Y^E\right)' V^{-1} \left(Y^M - Y^E\right),$$

where Y^M is a vector of outcomes generated by MINT, Y^E a vector generated by the external macroeconomic model, and V the covariance matrix of their difference. The vector of outcomes may represent, for example, various income components at a point in time, total retirement income at multiple future points in time, or a combination thereof. If $\left(Y^M - Y^E\right)$ is normally distributed, Q asymptotically follows the χ^2 distribution with a number of degrees equal to the dimension of the outcome vector.[33] Suppose that only one (scalar) outcome is compared to an external model. This special case reduces asymptotically to a simple z-test, with the variance given by:

$$V = \sigma^2_{Y^M} + \sigma^2_{Y^E} - 2\sigma_{Y^M Y^E}.$$

In practice, this formal statistical approach is unlikely to be feasible. The variance of Y^M is very difficult to determine, especially if MINT's projections need to be scaled up to reflect all cohorts; the covariance between Y^M and Y^E is unknown and probably not zero.

In practice, a more heuristic approach must therefore be taken. A "small" discrepancy, defined by the MINT user, may be acceptable, especially if differences in underlying assumptions and sample universes provide an explanation for the direction and magnitude of the discrepancy. A "large" discrepancy, however, should be analyzed carefully and may lead to the discovery of modeling errors.

Benchmarking

As a matter of principle, we do not advocate adjusting MINT model parameters to ensure a close match between aggregate MINT projections and external

---

[33] It may be preferable to formulate the test statistic in terms of the logarithm of outcomes.

macroeconomic projections. Discrepancies may arise from many sources, including from imperfections in macroeconomic models. However, we acknowledge the desirability of comparability of simulations conducted using MINT and other models, which is facilitated if the baseline simulations are equal.

If changes to MINT are contemplated to ensure a match of the aggregate projections of MINT and of an external macroeconomic model, we strongly recommend that every effort be made to preserve substantive conclusions from running MINT. For example, the calibration should not affect measures of income inequality such as the Gini coefficient. The following procedures may preserve MINT's conclusions.

- Adjust all weights proportionally. This may force equality of future income flows at the expense of a realistic current income flow.
- Adjust an intercept and/or a slope equally for all respondents. This may preserve current income flows and generate future flows that match those of an external macroeconomic model. While these types of adjustments ("fudge factors") are common in many macroeconomic models, we offer this option with great reluctance. The adjustment is arbitrary and the model is no longer capable of generating confidence intervals.

As a general rule, adjustments should be made equally for all simulants to prevent building in distributional effects that MINT itself does not generate.

4.3. Identify Demographic Models for Validating MINT

4.3.1. Introduction

This section identifies appropriate demographic forecasting models for validating, and possibly benchmarking, the MINT microsimulation model. The eventual objective is to support analyses of the extent to which aggregate forecasts implicit in MINT are consistent with external forecasts.

The process of validating and benchmarking a microsimulation model with demographic forecasts is exceedingly similar to the process with macroeconomic forecasts; thus, much of the discussion in Section 4.2 is applicable to validation with demographic forecasts. Rather than repeat criteria listed above, we consider this section as an extension of the previous one.

We proceed as follows. Section 4.3.2 outlines the ideal characteristics of a useful demographic forecasting model for validating MINT. We provide a brief review of demographic forecasting models and a discussion of the most promising models in Section 4.3.3. Further information about the theory and mechanics of aggregating, validating, and benchmarking MINT was discussed in Section 4.2 and will not be repeated here.

4.3.2. Demographic Model Characteristics

Before we turn to candidate demographic models, we first list the characteristics of a demographic model that would be appropriate for validating MINT. The ideal demographic model for evaluating MINT should:

- *Forecast through the year 2020*. MINT implicitly forecasts the population through the year 2020. A demographic model that serves to validate MINT projections should thus forecast through at least the year 2020.
- *Have been evaluated for accuracy.* The purpose of the validation procedure is to compare MINT to demographic forecasts that are considered "good." The external demographic model should therefore have a proven track record for accuracy. However, it is much less common to validate demographic models than it is to validate macroeconomic models. See Keyfitz (1981) for an exception.
- *Forecast population size and demographic characteristics by cohort.* MINT is specified to forecast retirement income of the 1931-1960 birth cohort only. Suitable demographic models should therefore forecast changes in the population by cohort, permitting a direct comparison.
- *Provide confidence intervals.* A formal statistical test of whether two forecasts differ requires confidence intervals (i.e., standard errors) for both forecasts.

Moreover, for policy analysis, standard errors are necessary to evaluate the potential risk under various scenarios.[34]

- *Be fully documented*. Projections of the demographic model and MINT may only be expected to be identical if their underlying assumptions are the same. Where identical assumptions are lacking (as will very often be the case), discrepancies may arise. An evaluation of the source of such discrepancies requires full knowledge of both models' underlying assumptions.

We limit our consideration to forecasting models that contain all these characteristics, with the exception of confidence intervals.[35]

4.3.3. Survey of Forecasting Models

Although there is long history of forecasting population trends, there are many fewer models in existence that are used for forecasting. We first provide a brief review of types of forecasting models that are generally used for demographic forecasting; then, we specifically discuss candidate models.[36, 37]

There are three primary categories of demographic forecasting methods. The first category is demographic accounting methods. These methods specify accounting identities for the underlying components of population growth. Then, these underlying components are forecasted into the future. For example, the population in a given year can be defined as the population in the previous year plus births last year and minus deaths last year (ignoring migration for this simple example). Births and deaths are forecast individually and population growth follows from the accounting identity. Usually, the underlying components are forecast using so-called "informed judgment methods." For example, the U.S. Census Bureau uses informed judgment methods. (See U.S. Bureau of the Census (1984) for a detailed description of its forecasts.) Although these models tend to provide reasonable long-term forecasts, standard errors are not available. To address the issue of variance, forecasts are sometimes made under "low," "intermediate," and "high" assumed levels for key

---

[34] Tuljapurkar (1992) provides a succinct introduction about why it is difficult to obtain standard errors with demographic forecasts and provides a discussion of the importance of standard errors for demographic forecasts in a decision theoretic framework.

[35] Additionally, one could require that the models are in the public domain, support projections under alternative policy regimes, and include behavioral feedbacks. These criteria, however, play a less important role for demographic models than they do for macroeconomic models. All demographic models that we encountered are in the public domain. Moreover, while changes to Social Security policy may theoretically have some effect on mortality, marriage, and divorce rates, these effects are likely to be small and perhaps ambiguous or controversial. In our assessment, these criteria do not merit further attention for purposes of evaluation and benchmarking MINT.

[36] Candidate models that are not reviewed here include DYNAMISM and MDM. Both models are older and do not focus on demographic forecasting. For more information on both models, see Gordon and Michel (1980) and the National Institute on Aging (1984).

[37] For a useful review of forecasting methodologies, see Land (1986). We borrow heavily from Land for this section. For a review that focuses on forecasting for older populations, see Guralnik, Yanagishita, and Schneider (1988).

trends, but the choice of such levels is typically arbitrary and the value of the additional projections difficult to assess. [38]

The second category is time series methods. These methods rely on projecting demographic quantities using standard time series models such as an AutoRegressive Integrated Moving Average (ARIMA) model. The models are purely statistical without underlying demographic justification. They tend to predict short- and medium-term population changes quite well.

The third category methods are structural in nature. Many macroeconomic models include endogenous population characteristics such as fertility, mortality, and marriage. The structural models are well equipped to forecast potential demographic changes to policy changes, but there is not a clear consensus that the added complexity is necessary to achieve sensible results for population forecasting.

Potential Demographic Forecasting Models

While there is a rich literature on models of mortality, marriage formation, and marriage dissolution, few models have been used to project population counts of the United States, by cohort and marital status. The best known and most widely cited population forecasts are those produced by the Bureau of the Census in its P25 Series of publications. Report P25-1130 projects population by age, sex, and race/ethnicity. The Census Bureau itself does not project population counts by marital status. In specialized reports, such as Bureau of the Census (1996, P23-190), population projections by marital status are provided. Those projections, however, are not produced by the Bureau of the Census, but by Felicitie Bell of the Office of the Chief Actuary of the Social Security Administration. The latest such report is Actuarial Study No. 112, Bell (1997).

Bell (1997) uses a demographic method based on accounting identities. Her projections include components for fertility, mortality, immigration, marriage, and divorce.[39] Following the practice of the U.S. Bureau of the Census, Bell uses informed judgment methods for each component series. Her assumptions, however, do not always correspond to those of the Bureau of the Census.

---

[38] Alho and Spencer (1990) provide a useful discussion of the difficulties of interpreting the high and low scenarios generated as part of the forecast.

[39] Bell (1997) projects for the Social Security Area, consisting of residents of the 50 states and the District of Columbia; armed forces overseas; civilian residents of Puerto Rico, the Virgin Islands, Guam, American Samoa, Palau, and the Northern Mariana Islands; Federal civilian employees overseas; dependents of Armed Forces and Federal employees overseas; crew members of merchant vessels; and other citizens overseas. The SIPP is representative of the civilian non-institutionalized population within the 50 states and the District of Columbia. Any comparison of MINT with Bell (1997) should thus adjust for military personnel, institutionalized persons, and persons outside the 50 states and the District of Columbia.

An important difference is in assumed reductions of mortality risks. Bell reports that the average annual percentage reductions in age-adjusted central death rates between 1900 and 1994 are 0.94 for men and 1.33 for women. In recent years, gains in longevity have been smaller: 0.78 percent annually for men between 1982 and 1994, and 0.54 percent for women. Bell' s intermediate assumption is that male and female mortality rates will both decrease at an average rate of 0.56 percent per year during the period 1994 through 2071. These reductions are far smaller than the average annual reduction observed between 1900 and 1994, about one-sixth below the reduction over the past twelve years, and in fact lower than any period of this century, with the exception of the 1954-1968 period. As a result of this assumption, life expectancy at birth is assumed to increase from 72.6 in 1995 to 77.5 in 2050 for males, and for females from 79.0 in 1995 to 82.9 in 2050.

The Census Bureau makes complicated adjustments to male and female age-specific death rates, resulting in life expectancy for males and females in 2050 of 79.7 and 84.3 years, respectively (P23-1130, Table B-1). The implicit reductions in mortality rates of the Census Bureau are thus substantially larger than those of Bell (1997).

Our MINT mortality projections assume annual reductions in mortality rates of 0.81 percent for males and 1.41 percent for females. These reductions are based on hazard model estimates of pooled Vital Statistics age-specific death rates from 1901 to 1994. Only mortality rates of individuals age 30 or over are taken into account in those estimates, as this is the relevant age group for MINT purposes.

Bell' s assumptions on marriage and divorce are based on trends in marriage and divorce rates over the past 15 years or so. Marriage rates are based on data from the Marriage Registration Area (MRA), which comprises 42 states and the District of Columbia, and covers approximately 80 percent of all marriages in the United States. As shown in Figure 4.1, marriage rates have declined mildly since about 1970. Bell (1997) assumed for the intermediate alternative that future age-adjusted rates of marriage for the Social Security Area would continue to slowly decrease and then stabilize in 2021. She inflated the number of MRA marriages to reflect the entire United States, but then reduced it by 5 percent to correct for Nevada; Nevada is not in the MRA, but has disproportionally many marriages.

Bell' s divorce rates are based on the Divorce Registration Area (DRA), which comprises 31 states and covers approximately 48 percent of all divorces. As shown in Figure 4.2, the divorce rate increased substantially in the 1960s and 1970s, but leveled off after 1980. Bell (1997) assumed under the intermediate alternative that throughout the projection period the age-adjusted divorce rate would remain close to the level as recently experienced.

MINT' s assumptions are based on hazard model estimates of marriage and divorce experiences reported by SIPP respondents in the Marital History Topical Module and subsequent panel reports until the end of the survey waves. For male marriage rates, we found a reduction of 0.79 percent annually; for females, we found a reduction of

0.36 percent annually (Table 2.5 on page 17). Male divorce rates were found to be flat since 1980; based on marriage histories reported by women, divorce rates have creeped up by 0.58 percent annually since 1980 (Table 2.7 on page 29).[40] We assume that those historical trends will continue through the projection period.

[40] Data limitations required estimating male and female divorce models separately, even though conceptually they follow the same process. See Section 2.4.

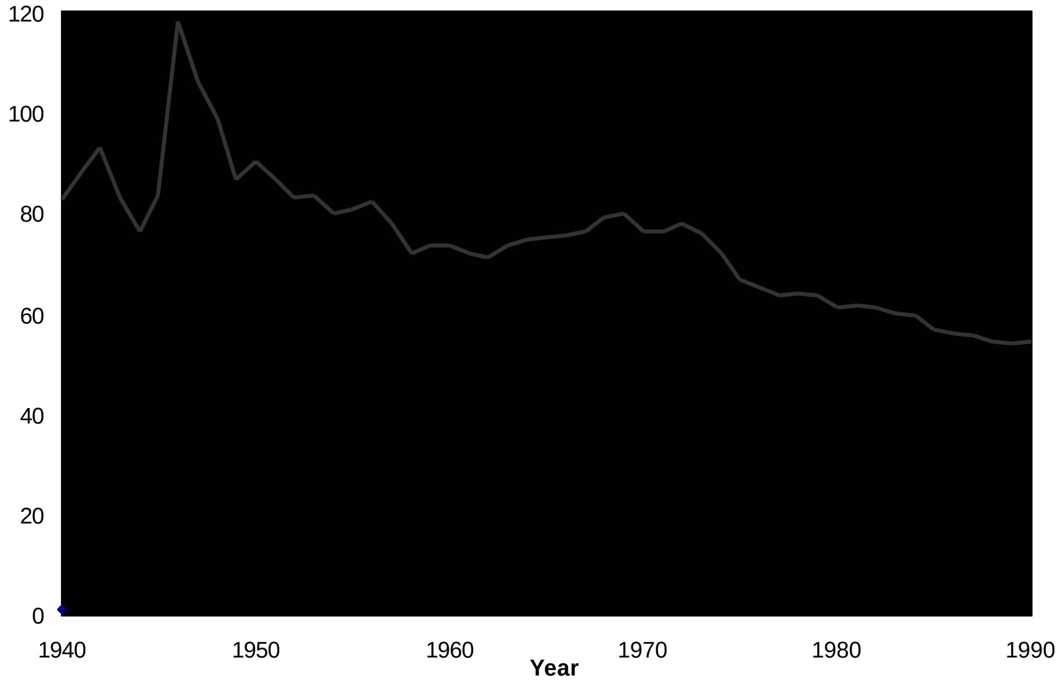

Figure 4.1. Marriages per 1,000 Unmarried Women Aged 15+, 1940-1990

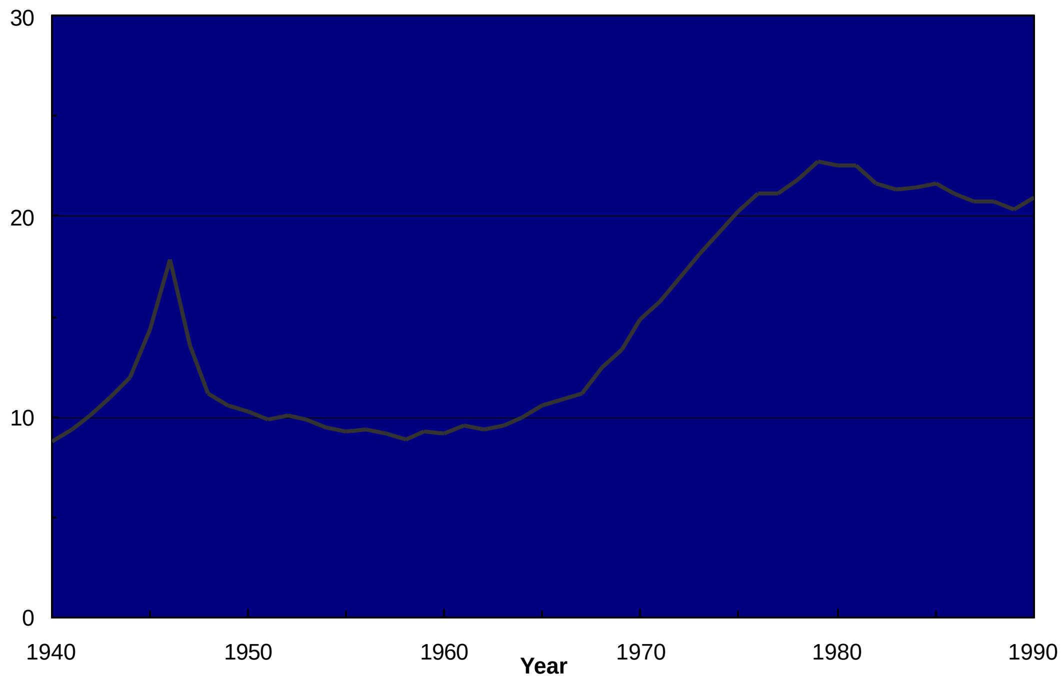

Figure 4.2. Divorce Rate per 1,000 Married Women Aged 15+, 1940-1990

Section 4.5 compares MINT's projections by marital status to Bell (1997).

Both Bell (1997) and the Bureau of the Census (1996b) use "informed judgment methods" to obtain assumptions on future trends in demographic submodels. Both also formulate "low-cost," "intermediate," and "high-cost" scenarios. A drawback to this approach is that standard errors are not available for the forecasts and that the assumed high/low ranges for births, deaths, and other demographic components are not probabilistically consistent with one another. Furthermore, the choice of low-cost and high-cost scenarios is largely arbitrary and their corresponding projections difficult to interpret.

Ronald Lee and Shripad Tuljapurkar address this issue in a series of papers. See, for example, Lee and Tuljapurkar (1994). The Lee-Tuljapurkar approach uses a combination of demographic models of fertility and mortality and a statistical time series model. In the first step, simple demographic models of age- and time-specific fertility and mortality are estimated, resulting in estimates of time trends with standard errors (much like our estimates of the mortality, marriage formation, and divorce time trends from Vital Statistics and SIPP). The estimates are translated into stochastic transition matrices. In the second step, most recent population counts (by sex and age) are multiplied by the transition matrices to obtain next-period population counts (with a correction for assumed immigration). The number of projection period years determines the number of matrix multiplications. This stochastic method enables Lee and Tuljapurkar to generate standard errors and confidence bands for their forecasts.

The Lee and Tuljapurkar (1994) expected total population projection for the year 2065 corresponds closely to the Census Bureau's intermediate projection, and their 95 percent confidence intervals are close to the low and high scenario projections. However, Census low/high intervals for elderly subpopulations are much wider than those found using stochastic forecasts, whereas the opposite was found for youth and elderly dependency ratios. [41]

We believe that the Lee-Tuljapurkar method offers a far superior alternative to the low/high scenarios used by Bell (1997), the Bureau of the Census (1996b), and many other "informed judgment" approaches. It is internally consistent, statistically sound, and not subject to arbitrary assumptions (except on exogenous influences, such as immigration).

Recommendation

An attractive feature of Bell (1997) is that it projects population counts by marital status. However, we question Bell's intermediate assumption of 0.56 percent annual

[41] Alho and Spencer (1990) attempted to relate high/low forecasts of the Office of the Actuary to confidence intervals. They found that the intervals may not be interpreted as confidence intervals, and that the high/low intervals tend to be wider than the 95 percent confidence bands.

mortality rate reductions. Comparisons of MINT population forecasts with Bell (1997) result in larger MINT projections because of MINT's assumption of more rapid gains in longevity; see Section 4.4. Additional discrepancies may result from differences in the population under consideration. In particular, Bell's inclusion of the military and the incarcerated or otherwise institutionalized population is likely to exacerbate the discrepancy. [42]

Projections of the Bureau of the Census (1996b, Current Population Series P23-1130) and Lee-Tuljapurkar (1994) yield aggregate results that are very close to each other. The Census low/high intervals, however, differ from Lee-Tuljapurkar's 95 percent confidence bands. Given the arbitrary nature of the Census Bureau's high/low assumptions and the internal consistency of the Lee-Tuljapurkar method, we favor benchmarking MINT's aggregate population projections with those of Lee and Tuljapurkar (1994).

Lee and Tuljapurkar (1994) project expected aggregate population size in 2020 of 316 million; the number of individuals age 65+ is projected at 52 million. MINT is restricted to the 1931-1960 birth cohort, excludes immigrants, the military, incarcerated and other institutionalized individuals, and should thus project smaller population counts.

Neither Lee and Tuljapurkar (1994) nor the Bureau of the Census (1996b) provide estimates of population counts by marital status. We therefore recommend that MINT projections by marital status be benchmarked against the results in Bell (1997). As argued above, we question Bell's mortality assumptions, but her projected shares of individuals that are never married, married, widowed, and divorced may serve as useful benchmarks. Table 4.3 shows the projected fractions by age group for the year 2020.[43]

Table 4.3. Projected Marital Status of Persons Age 65 and Over by Sex, 2020 (Bureau of the Census 1996a)

	Males				Females			
	Single	Married	Widowed	Divorced	Single	Married	Widowed	Divorced
Age 65+	6.2	72.1	12.7	8.9	5.0	43.6	37.1	14.3
Age 65-74	7.5	75.0	7.6	9.9	5.6	55.4	22.4	16.6
Age 75+	4.0	66.9	22.1	7.0	4.3	28.3	56.2	11.2

[42] Bell's high-cost alternative assumes mortality rate reductions about the same as for 1900 through 1994. That assumption, however, is combined with high-cost marriage and divorce rates.

[43] These figures are based on Bureau of the Census (1996a), Table 6-1, and provide greater detail than Bell (1997).

4.4. MINT vs SSA OACT Longevity Projections

Our mortality projections are based on hazard model estimates from PSID data that are corrected such that, in the aggregate, mortality rates are identical to those based on Vital Statistics data. See Section 2.2.3. Vital Statistics from 1901 through 1994 indicate that the log-hazard of mortality has decreased at an annual rate of 0.81 percent (males) and 1.41 percent (females). A key assumption underlying our mortality projections is that mortality rates will continue to improve at this pace.

Table 4.4. Historical Average Annual Percentage Reductions in Age-Adjusted Central Death Rates (Bell, 1997)

	1900-36	1936-54	1954-68	1968-82	1982-94	1900-94
Male						
0-14	2.91	4.75	1.66	4.39	2.60	3.26
15-64	1.02	1.91	-.20	2.22	.61	1.14
65-84	.20	1.15	-.13	1.47	1.21	.65
85+	.22	1.21	-.89	1.56	-.34	.38
65+	.20	1.16	-.33	1.49	.79	.58
Total	.78	1.60	-.21	1.78	.78	.94
Female						
0-14	3.12	5.01	1.72	4.19	2.49	3.36
15-64	1.19	3.62	.57	2.20	.70	1.66
65-84	.36	2.06	1.07	2.01	.58	1.07
85+	.23	1.21	.13	2.06	.09	.66
65+	.32	1.82	.77	2.03	.42	.95
Total	.90	2.47	.77	2.15	.54	1.33

Based on Table 4.4, SSA OACT observes the following (Bell 1997):

> An examination of the age-adjusted death rates since 1900 reveals several distinct periods of mortality reduction. During the period 1900 to 1936, annual mortality reduction averaged about 0.8 percent for males and 0.9 percent for females. Following this was a period of rapid reduction, 1936 to 1954, in which mortality decreased an average of 1.6 percent per year for males and 2.5 percent for females. The period 1954 to 1968 saw an actual increase for males of 0.2 percent per year and a much slower reduction of 0.8 percent per year for females. From 1968 through 1982 rapid reduction in mortality resumed, averaging 1.8 percent for males and 2.2 percent for females, annually. From 1982 to 1994, slower reduction in mortality resumed, decreasing an average of 0.8 percent for males and 0.5 percent for females.

After reviewing cause-specific mortality rates, Bell (1997) makes the following assumption:

> After adjustment for changes in the age and sex distribution of the population, the intermediate alternative mortality is projected to decrease at an average rate of 0.56 percent per year during the period 1994 through 2071, about half the average annual reduction observed during 1900 through 1994, but greater than the female rate of reduction for the 1982 through 1994 period.

As may be expected, our assumed 0.81 percent (males) and 1.41 percent (females) annual decrease in mortality rates implies greater projected gains in longevity than those based on the 0.56 percent (males and females) assumed by SSA OACT.

The first row in Table 4.5 shows remaining life expectancies for a 65-year-old person, by sex and year, as generated by the projection algorithms of OACT and MINT.

Table 4.5. Remaining Life Expectancies at Age 65 by Sex and Year

	Male		Female	
Year	OACT	MINT	OACT	MINT
1995	15.6	15.2	19.0	19.5
2005	16.0	15.8	19.5	20.6
2015	16.4	16.5	19.8	21.8
2025	16.8	17.1	20.2	22.9
2035	17.3	17.7	20.7	24.1

The 1995 OACT figures are actual figures from Vital Statistics. Note that MINT implies a slightly lower life expectancy for men in 1995 and a slightly higher one for females. This is because MINT' s mortality rates follow from a fitted model; the 1995 actual life expectancy for men was above the trend, while for women, it was below the trend.

As expected, MINT projects faster gains in longevity than OACT. Between 1995 and 2035, MINT projects gain for men of 2.5 years, while OACT projects gains of only 1.7 years. For women, the difference is larger: MINT 4.6 years vs. OACT 1.7 years.

It should be noted that MINT distinguishes a male and a female time trend, and it assumes that the 1901-1994 trends will continue. These trends imply a continued and further widening of female longevity advantage. Over the recent past, we find no evidence of male-female convergence, and we therefore project a continued divergence for the next twenty-five years or so. In the long run, however, male and female trends may converge or not diverge further. We are therefore hesitant to project much beyond the MINT horizon of 2020.

4.4.1. "Current" and "Cohort" Life Expectancies

Publications such as the *Vital Statistics of the United States* series (e.g., National Center for Health Statistics, 1998) contain so-called current, "snapshot," or "cross-sectional" lifetables, which report age-specific mortality rates of the population over the period of interest. The many mortality rates in such lifetables are often summarized in a life expectancy figure. This life expectancy is based on a synthetic cohort of individuals and its computation assumes that this cohort is subject throughout its existence to the age-specific mortality rates observed for an actual population during a particular period. For example, the 1995 life expectancy at birth (75.8 years) assumes that someone who is born in 1995 will face the same mortality risks at age ten as a ten-year-old in 1995, and the same mortality risks at age 60 as a 60-year-old in 1995, etc. The MINT life expectancies reported in Table 4.5 are computed using projected "current" mortality rates as of the years indicated in the first column. While not explicitly stated, the OACT figures are also based on "current" mortality rates; the 1995 life expectancies are virtually identical to those reported in *Vital Statistics of the United States, 1995* (National Center for Health Statistics, 1998).

However, longevity has steadily increased over the past century as a result of improved nutrition, health habits, medical technology, etc. If such improvements continue in the future, a 65-year-old in 2060 is likely to experience far more favorable survival chances than a 65-year-old in 1995. The average lifespan of all children born in 1995 is thus likely to be greater than the current life expectancy of 75.8 years. How much greater depends on the rate at which mortality risks decrease.

Given the important role that trends in mortality risks play in MINT projections, we constructed "cohort" lifetables and life expectancies, also known as "longitudinal" and "generational" lifetables (National Center for Health Statistics, 1997). These cohort life expectancies are based on projected mortality rates, taking into account lower future mortality rates, as projected by the model. In other words, the remaining life expectancy for a 65-year-old in 1995 is computed using the mortality rate that was actually experienced by 65-year-olds in 1995,[44] the mortality rate that is projected for a 66-year-old in 1996, the rate projected for a 67-year-old in 1997, the rate projected for a 68-year-old in 1998, etc.

Figure 4.3 shows a stylized model of cohort mortality rates that is not based on any empirical estimates; it only serves to illustrate the difference between current and cohort rates. The top line represents the hypothetical mortality rate (in logarithmic form) of the population in 1995; slightly below it is the hypothetical pattern of the 1996 population; through the 2015 population. As shown in the figure, mortality rates are assumed to decrease steadily over time. Consistent with our mortality model, all mortality rates decrease with the same annual percentage reduction, i.e.,

[44] The MINT mortality model is based on 1901-1994 Vital Statistics, so strictly speaking, the 1995 rate for 65-year-olds is a projection.

the mortality patterns are parallel in logarithmic form. Following the 65-year-old in 1995 to older age and future years, we note that the effective mortality risks that he will experience (denoted by the darker line) increase less steeply with age than those of any current mortality risk pattern.

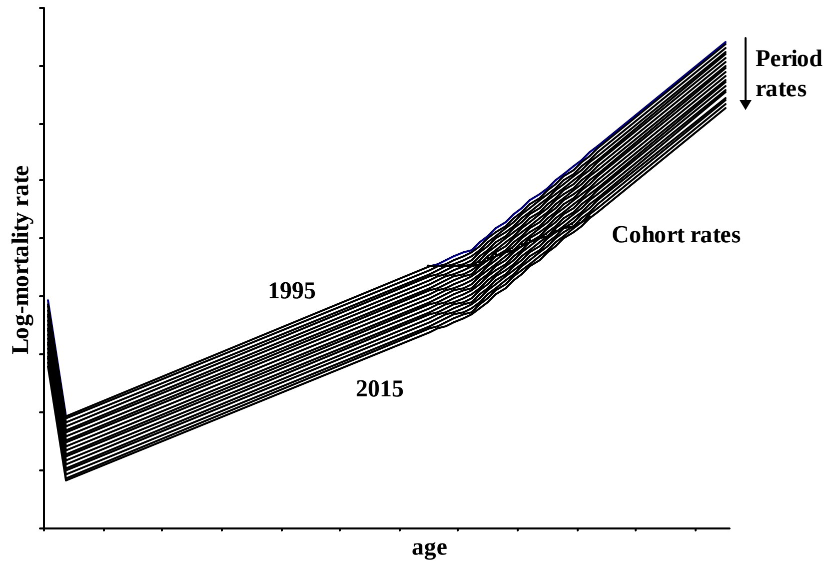

Figure 4.3. Stylized Model of Cohort Mortality Rates

The difference in current and cohort life expectancies is substantial, as Table 4.6 shows. In 1995, 65-year-old men may expect to live about 0.8 years longer than standard, current lifetables indicate. The difference is even larger for women, who may expect to benefit about 1.8 years from future developments that prolong life. The difference is larger for women because of their faster mortality rate reductions: 1.41 percent annually compared to 0.81 percent annually for men.

Table 4.6. Current and Cohort Remaining Life Expectancies at Age 65 (MINT)

	Male		Female	
Year	Current	Cohort	Current	Cohort
1995	15.2	16.0	19.5	21.3
2005	15.8	16.6	20.6	22.6
2015	16.5	17.3	21.8	23.9
2025	17.1	18.0	22.9	25.2
2035	17.7	18.7	24.1	26.6

For comparisons of life expectancies projected by MINT and external demographic models, any consistently defined life expectancy measure will suffice. Given the practice of SSA's OACT (and the Bureau of the Census, 1996a, 1996b), we presented such comparisons using current life expectancies (Table 4.5 above). However, such life expectancies substantially underestimate the number of years that an individual may expect to live, with potentially serious implications for the timing of retirement, savings behavior, etc. For such purposes, cohort lifetables and cohort life expectancies are superior. The Urban Institute converted wealth stocks into retirement income flows using an annuitization algorithm that RAND developed (Toder et al., 1999, pages 62 and 201); that algorithm uses cohort mortality rates.

4.5. MINT vs. SSA OACT Projections by Marital Status and Sex

We now turn to a comparison of MINT and OACT population distribution forecasts by marital status and sex. The OACT forecast method is documented in Bell (1997), *Social Security Area Population Projections: 1997*. That publication, however, does not contain full details by age group. We therefore also rely on Table 6-1 of Bureau of the Census (1996a), *65+ in the United States* (Current Population Reports, P23-190). That Census Bureau report is based on OACT' s *Social Security Area Population Projections.*

As discussed earlier, OACT assumes a 0.56 percent annual mortality reduction and thus projects a smaller population size than MINT. For purposes of the current comparison, we ignore the number of deceased persons and focus on the distribution by sex and marital status only. The comparison refers to January 1, 2020. OACT projections cover all birth cohorts; MINT projections only cover individuals born in 1931-1960 birth cohort (59-88-year-olds in 2020).

OACT generates three projections: an intermediate, low-cost, and high-cost scenario. The low-cost and high-cost assumptions are chosen by consensus opinion about the plausible ranges of the forecasted series. They lack statistical basis and may not be interpreted as confidence intervals. For comparison purposes, we therefore restrict ourselves to the intermediate forecasts.

Table 4.7. Demographic Distribution in 2020, ages 65+ (percent)

	Males		Females	
	MINT	OACT	MINT	OACT
Never married	4.7	6.2	5.6	5.0
Married	76.3	72.1	46.3	43.6
Widowed	6.5	12.7	29.5	37.1
Divorced	12.5	8.9	18.6	14.3
Total	100.0	100.0	100.0	100.0

Table 4.7 shows the distribution of marital status by sex as generated by MINT and by OACT' s intermediate forecasts. Several discrepancies deserve attention. First, MINT projects a lower fraction of never married men. This may in part be attributed to MINT' s mortality model, which accounts for differential survival of never married males. As Table 2.1 on page 17 indicates, never married males experience mortality rates that are about 21 percent higher than those experienced by married men. The resulting shorter life expectancy implies that disproportionally many never married men will have become deceased by 2020.

Second, OACT projects higher widowhood rates in 2020 than MINT. This may in part be attributable to OACT' s conservative assumption about future gains in longevity. As a result, OACT projects higher mortality and thus higher widowhood rates. Another factor is the full 65+ age range covered by OACT projections; MINT projections only apply to individuals up to age 88 in 2020 (the 1931 birth cohort). The 89+ population contains disproportionally many widows, as shown in Figure 2.10.

Third, MINT projects somewhat higher fractions of married individuals, a necessary implication of lower mortality and widowhood rates. The higher projected number of married couples implies higher Social Security expenses on spousal benefits.

Fourth, MINT projects higher fractions of divorced individuals. As shown in Section 2.8, about 61 percent of divorced women at age 62 were married more than ten years and thus potentially eligible for benefits on the basis of their ex-husband's earnings. MINT may thus project greater outlays on spousal benefits than OACT, depending on the assumptions OACT makes about the fraction divorcées that receives spousal benefits.

Table 4.8. Demographic Distribution in 2020, ages 75+ (percent)

	Males		Females	
	MINT	OACT	MINT	OACT
Never married	3.5	4.0	4.3	4.3
Married	78.4	66.9	37.2	28.3
Widowed	7.4	22.1	42.0	56.2
Divorced	10.8	7.0	16.5	11.2
Total	100.0	100.0	100.0	100.0

Table 4.8 shows the distribution of individuals age 75 and older in 2020, by marital status and sex. Elderly men are projected to be predominantly married. However, their numbers have thinned, resulting in an increased fraction of widowed women. The discrepancies between MINT and OACT projections of individuals age 65 and older (Table 4.7) persist and become more pronounced at ages 75 and older in Table 4.8.

5. Individual Income Tax Model

5.1. Introduction

MINT' s income projections, as produced by The Urban Institute, represent pre-tax income flows through the year 2031 (Toder et al., 1999). SSA' s DPE wants to have the ability to project after-tax poverty rates and to evaluate after-tax consequences of reform proposals. To that end, RAND developed an individual income tax model.[45]

This chapter documents this income tax model. Also see Klerman and Panis (1999). The model consists of a SAS macro that approximates the federal and state taxes corresponding to the profile of income provided by the main SSA MINT model. We begin by describing how to use the macro. We then discuss the assumptions behind the model and how to modify if (or when) the tax code changes.

[45] We gratefully acknowledge substantial expert input from Jacob Klerman in the development of the tax model.

5.2. Model Input and Output

The taxation model consists mainly of a SAS macro, `%computax` It takes two formal input arguments and returns three output arguments. It assumes the existence of many (income, marriage, and demographic) variables, and requires that several arrays and formats have been declared—see below.

This macro takes as formal input:

- `year`: the year (four digits) for which taxes need to be computed. This argument may be a number (such as 2020) or a variable name (such as year);
- `assetinc` the name of the *array* containing asset income variables. Needed to allow for asset income on the basis of multivariate as well as unisex lifetables.

and provides as output:

- `fedtax` federal income taxes. This argument must be a variable name;
- `ficatax` Federal Insurance Contributions Act (FICA) taxes, i.e., the sum of Old-Age, Survivors, Disability (OASDI), and Hospital Insurance (HI) taxes. This argument must be a variable name;
- `statetax` total sub-federal taxes, including state income and sales taxes and sub-state taxes (e.g., county and local income, sales, property, and use taxes). This argument must be a variable name.

Each of these three tax variables corresponds to the year and profile of income as defined in the current observation of the MINT data.

5.2.1. A Simple Example

The tax model may be best illustrated using a sample program. Assume that the MINT data set is `mint.sd2`and located in the same directory as the following program:[46]

```
1     libname in  '.';
2
3     %include 'cpi.sas';      /* CPI series      */
4     %include 'ssawage.sas';  /* SSA wage series */
5     %include 'marstat.sas';  /* macro to figure out marital status */
6     %include 'computax.sas'; /* macro to compute taxes */
7
8     data new;
```

[46] The line numbers on the far left of the listing are not part of the code. They are included to ease description of the code. They do not appear in the actual source code.

```
 9       set in.mint;
10
11       array howend(*)  howend1-howend12;
12       array marb(*)    marb_1-marb_12;
13       array mare(*)    mare_1-mare_12;
14       array spbdate(*) spbdat1-spbdat12;
15
16       array inde(1990:2031) inde1990-inde2031;
17       array sern(1990:2031) sern1990-sern2031;
18       array hpen(1990:2031) hpen1990-hpen2031;
19       array spen(1990:2031) spen1990-spen2031;
20       array inci(1990:2031) inci1990-inci2031;
21       array incu(1990:2031) incu1990-incu2031;
22       array ssb(1990:2031)  ssb1990-ssb2031;
23       array sssb(1990:2031) sssb1990-sssb2031;
24
25       /* Compute the tax liability for the year 2020 using  */
26       /* asset income from multivariate lifetables:         */
27       %computax(2020,inci,federal,fica,state);
28
29    run;
```

Line 1 identifies the location of the data and programs.

Lines 3-6 include external files into the program. The first file, `cpi.sas`, contains SAS code with a proc format that defines format `cpi`. This format is used to conveniently map years into the corresponding (projected) Consumer Price Index for Urban Wage Earners and Clerical Workers (CPI). This format is required by the tax model macro. See Appendix B.1 for a listing of `cpi.sas`.

The second included file, `ssawage.sas`, contains SAS code with a `proc format` that defines format `ssawage`. This format is used to conveniently map years into the corresponding (projected) Social Security average wage. All monetary amounts in the MINT income and asset projections are relative to the Social Security average wage. This format is required by the tax model macro. See Appendix B.2 for a listing of `ssawage.sas`.

The third included file, `marstat.sas`, contains SAS code which defines macro `marstat`. This macro is required by the tax model macro to determine an individual's marital status as of the end of the tax year. See Appendix B.3 for a listing of `marstat.sas`.

The fourth included file, `computax.sas`, contains SAS code which defines macro `computax`. This macro computes tax liabilities and is the main component of the MINT tax model. See Appendix B.4 for a listing of `computax.sas`.

Lines 11-14 declare arrays which are required by `marstat` and `computax`. They presume the existence in `mint.ssd01` of the following variables:

- `howend1-howend12`: disposition of marriages;

- marb_1-marb_12 wedding dates;
- mare_1-mare_12 end dates of marriages;
- spbdat1-spbdat12 birth dates of spouses.

Lines 16-23 declare additional arrays which are required by marstat and computax. They presume the existence in mint.ssd01 of the following variables:

- inde1990-inde2031 respondent earnings
- sern1990-sern2031 spousal earnings
- hpen1990-hpen2031 respondent defined benefit pensions
- spen1990-spen2031 spousal defined benefit pensions
- inci1990-inci2031 annuitized asset income (multivariate lifetables)
- incu1990-incu2031 annuitized asset income (unisex lifetables)
- ssb1990-ssb2031 respondent Social Security benefits
- sssb1990-sssb2031 spousal Social Security benefits

Line 17 calls macro computax, which computes tax liabilities. In this example, liabilities for calendar year 2020 are computed using asset income from multivariate lifetables. Three new variables are created, corresponding to federal income tax (federal), FICA taxes (fica), and state and local total taxes (state).

Notes:
1. The year input argument may be either a number (such as 2020) or a variable name. Only years 1990 through 2031 are supported.
2. The asset income argument must be the name of an array, as defined by the user.
3. The three tax liability output arguments must be variable names, chosen by the user.
4. To compute tax liabilities of income flows including asset income based on unisex lifetables, specify incu as the second argument.

5.3. Model Assumptions

Income tax laws are very complicated and many potentially relevant details are unknown to the MINT user, especially in future years. We therefore make a number of simplifying assumptions. The most important assumptions are:

- Respondents that are unmarried as of the end of the year file a single tax return; married respondents file a joint tax return. There is one exception: individuals who have become widowed during the reference year and who did not remarry in that year file as married.
- There are no dependent children for whom an exemption may be claimed (line 6).
- There is no income from unemployment compensation (line 12).
- No deductible IRA contributions are made (line 15).
- No student loan interest deduction may be made (line 16).
- Respondents take the standard deduction, i.e., do not itemize expenses. There are no deductible medical savings account contributions, moving expenses, penalties on early withdrawal of savings, alimony expenses, or other expenses which affect Adjusted Gross Income (AGI). The standard deduction takes account of the respondent’ s (and spousal) age, but we assume that he/she is not blind.
- Respondents are not eligible for tax credits due to disability, child care, education, adoption, foreign tax payments, or other factors. They may, however, be eligible for the tax credit for the elderly (line 27).
- To the extent that there is earned income, we assume that the taxpayer is an employee, i.e., FICA taxes do not include the employer portion. Also, there are no deductible self-employment taxes and no contributions to Keogh or other self-employed Defined Contribution (DC) pension plans.
- Each tax year may be considered independently of other years, i.e., there is no carry-over of income across fiscal years.

The MINT simulation data contain information on projected income flows from four major categories: earned income, defined benefit (DB) pension income, Social Security benefits; and income from assets, including defined contribution (DC) pension balances. It also contains aggregate income projections (the sum of income components), but such income flows are only computed for the period after the respondents are projected to become entitled to Social Security benefits. The tax model applies to all years, including those before entitlement for Social Security benefits. The model is therefore solely based on projections of income components; aggregate income variables are not utilized in the computations.

The tax model is based on 1998 tax laws. With two exceptions, we assume that fiscal amounts (thresholds, standard deductions, exemptions) will continue to be adjusted according to the Consumer Price Index for Urban Wage Earners and Clerical Workers (CPI). The first exception is the Social Security Contribution Base, above which no OASDI contributions are made. This base, $68,400 in 1998, is assumed to increase in proportion with projections in the Social Security average wage index. The second exception relates to the computation of the taxable Social Security benefits. In 1998, up to half of Social Security benefits of a married couple is taxable if total income (defined according to the rules specified in the law) exceeds $32,000, and up to 85 percent is taxable if total income exceeds $44,000. The thresholds for single individuals are $25,000 and $34,000, respectively. These thresholds are not indexed and will thus remain the same throughout future years.

For years prior to 1998, the model assumes that the 1998 laws apply, with discounted monetary amounts in accordance with the CPI and Social Security average wage index. This may lead to small discrepancies with actual past tax liabilities.

The tax model only supports calculations for years 1990 through 2031. The lower bound was chosen because MINT data are based on 1990-93 SIPP panels, so that there are no income projections prior to 1990. The upper bound was chosen as the last year for which MINT income projections are available. Any attempt to compute taxes outside the supported range results in abortion of the SAS program, with an informative error message.

5.3.1. Income from Assets

Income from assets is approximated in MINT as the annuity amount a family could purchase if it annuitized 80 percent of its financial assets. Two annuity flows are available. The first (variables incu1980-incu2031) are based on ‘ unisex’ lifetables which only account for differential remaining longevity by age, using 1990 lifetables; the second (variables inci1980-inci2031) are based on ‘ multivariate’ lifetables which account for differential longevity by age, sex, race, education, and calendar time.

Assets include both tax-sheltered wealth (IRAs, Keoghs, and DC pension wealth), and after-tax savings. Any cash flow from tax-sheltered assets, whether in the form of withdrawal or interest/dividend, is taxable. Interest and dividend from after-tax savings are taxable as ordinary income; withdrawals of the principle may be subject to capital gains taxation.

The information in the MINT data is insufficient to determine how much income from assets is taxable. The tax model therefore adopts a very crude rule of thumb: a constant fraction of income from assets is taxable. That fraction is modeled as a *user-modifiable parameter*, `gamma` See source code line 23 in `computax.sas` (Appendix B.4). As directed by the SSA Task Manager, gamma is presently set to 1, i.e., all income from assets is assumed to be taxable. As stated above, note that the income flows are based on annuitization of 80 percent of assets.

5.3.2. State and Local Total Taxation for the Elderly

MINT projects future income flows and demographic status. It does not project future state of residence. It is therefore impossible to compute state income tax liabilities.

State income tax regimes vary widely. Nine states do not levy personal income tax at all; 25 states and the District of Columbia base state income tax on federal AGI; eight states base tax liabilities on federal taxable income; two states base state income tax on federal income tax liability; the remainder specify their own tax basis. Furthermore, states vary in the treatment of public and private pensions, with 25 states fully or partially exempting public pensions, and 36 states fully or partially exempting private pensions. They also differ in tax rates. The top marginal tax rate among states that levy personal income taxes varies from 2.8 percent in Pennsylvania to 12.0 percent in North Dakota. Thus, given any income level, the state income tax burden varies widely.

However, when considering the total state and local tax and fee burden, the differences are much smaller. As shown by Kroes (1998), excluding Alaska, the total state and local tax burden in 1994-95 ranges from 11.0 percent of personal income in New Hampshire to 17.2 percent in New York. The median state is California at 14.1 percent. Alaska—a state that does not levy state personal income tax—stands out with 23.6 percent. In other words, while there is substantial variation across states in personal income tax burden, state and local legislatures tend to compensate through higher or lower county and city taxes, and through various fees.

Against this background, the SSA Task Manager decided that the tax model approximate state and local total tax burden as a constant fraction of federal income tax liability. Given that California is at the median state and local tax burden, and a large state, the default applicable flat percentage rate is the ratio of California taxes (including local taxes and fees) to federal taxes for the elderly population. The fraction is modeled as a *user-modifiable parameter*, `lambda`. See source code line 20 in `computax.sas` (Appendix B.4). We estimate this ratio for California to be 0.835. At present, `lambda` is thus set to 0.835, i.e., the `computax` macro returns `statetax` as 83.5 percent of `fedtax`. The remainder of this subsection explains how we derived this estimate.

An exact figure for the percentage of California state and local fees per $1,000 personal income among the elderly is not available. Our approximation is based on estimates of tax and fee components and the fraction of income among the elderly which is taxable. Table 5.1 shows the sources of California tax and fees revenues as estimated by Kroes (1997).

Table 5.1. California State and Local Taxes and Fees (1994-1995; per $1,000 Personal Income)

Fees and assessments	$32.51
General sales tax	29.19
Personal income tax	24.61
Corporate income tax	7.71
Property tax	30.25
Total[a]	141.44

[a] Note: the sum of components does not add up to the total. We contacted the author but did not receive a response.

We assume that fees and assessments, general sales tax and property taxes per $1,000 personal income are roughly the same for the elderly and general population. Corporate income tax is not levied on personal income and is thus not applicable. The personal income tax figure needs to be adjusted downward because not all income among the elderly is subject to California state income taxation. Social Security benefits account for 42 percent of all income for persons age 65 and over, while pension income and annuities account for 19 percent (Baer, 1997). California has a broad-based personal income tax exemption of Social Security benefits, but it allows no exemptions for pensions or other retirement income that is counted in federal AGI. In other words, approximately 58 percent of income among the elderly is subject to California state income taxation. Ignoring progressivity effects, we therefore estimate the total state and local tax and fee burden at approximately $106.22 (= 32.51 + 29.19 + 0.58*24.61 + 30.25) per $1,000 personal income.

We apply the same adjustments to federal income taxation. Kroes (1997) estimates that the federal tax burden for Californians is $219.20 per $1,000 personal income. Ignoring partial taxation of Social Security benefits, the average federal income tax for Californians age 65 and over is $127.14 (= 0.58*219.20) per $1,000 personal income.

For Californians age 65 and over, the ratio of California state and local taxes and fees to federal taxes is thus 0.835 (= 106.22/127.14).

5.4. Technical Notes

The model is built on the 1998 Federal Income Tax forms, in particular form 1040A and supporting forms, worksheets, and schedules.[47] To a great extent the internals of the model preserve the logic, computational approach, and variable names corresponding to those forms and schedules.[48] Anyone wishing to understand or modify the macro is urged to have copies of the relevant forms at hand. To that end, those forms are reproduced at the end of this guide. Appendix B.5 contains 1998 Form 1940A; Appendix B.6 contains the Social Security Benefits Worksheet; and Appendix B.7 contains Form 1040A Schedule 3.

Taxable Social Security benefits are computed according to the 1040A worksheet for lines 13a and 13b (see the 1040A Instructions at page 27). We note that the income cutoffs here are the only place where dollar amounts on tax forms are not indexed by the assumed CPI.

Filing status is determined by a simple rule. Married couples are assumed to file jointly. Single individuals are assumed to file as "Single." The return is assumed to involve no dependents beyond the head and possibly a spouse. The number of deductions is computed according to the worksheet on the 1998 1040A Instructions at page 31 (see the macro for lines A20a and A21).

Taxable income (A24) is then simply computed as AGI less the value of the exemptions and the deductions (properly computed for the age of the head and spouse). Given this computed value for Taxable Income (and its assumptions), federal taxes are then computed using the formulas provided with the 1040 instructions. They are within rounding error of the values in the 1040A tax tables. Using the formulas results in macro which is shorter and easier to maintain (i.e., update when the tax law changes). In particular, the macro proceeds using the break-points for the brackets and the tax rates for income within the brackets.

The macro then computes FICA (OASDI and HI) taxes. The sum of these two taxes is returned in the variable `ficatax`(it would be simple to break out the two taxes if desired). The computations are based on earned income. We note that FICA is the only place where separate income for head and spouse is required to compute taxes.

Finally, the macro computes an approximation to total state tax payments, `statetax`, as user-modifiable parameter `lambda`times federal income tax liability,

[47] We opted for the structure of Form 1040A, rather than 1040, because the assumptions stated above rule out any complication for which Form 1040 would be required. We relax the Form 1040A restriction that taxable income must be less than $50,000 by using tax rate schedules, rather than tax tables.

[48] Variables prefixed with an "A" correspond to 1998 Form 1040A line numbers; those prefixed with a "W" correspond to Social Security Benefits Worksheet line numbers; those prefixed with a "C" correspond to line numbers on Schedule 3.

`fedtax` This fraction is intended to include all sub-federal taxes, including state income taxes, state sales taxes, and sub-state (local; i.e. county, city, etc.) taxes.

5.5. Customization

The tax model may be readily customized to support alternative assumptions on (future) tax regimes. We highlight three aspects and illustrate modifications.

5.5.1. Taxation of Income from Assets

As explained above, MINT estimates income from assets as the annuity flow that a family could purchase if it annuitized 80 percent of its financial assets. Since insufficient information is available to determine how much income from assets is taxable, the tax model adopts a very crude rule of thumb: a constant fraction of income from assets is taxable. That fraction is modeled as a user-modifiable parameter, `gamma`. This parameter is currently set to one, i.e., all income from assets is assumed to be taxable (source code line 23 in `computax.sas`):

```
%let gamma=1;
```

This parameter may be modified by the user. For example, to assume that 75 percent of asset income is taxable, change line 23 to:

```
%let gamma=0.75;
```

5.5.2. State and Local Total Taxation

As explained above, the tax model approximates state and local total tax burden—including state and local personal income tax, sales tax, property tax, and fees—as a constant fraction of federal income tax liability. The fraction is specified as a user-modifiable parameter, `lambda`. This parameter is currently set to 0.835, i.e., the `computax` macro returns `statetax` as 83.5 percent of `fedtax` (line 20 of the `computax` macro):

```
%let lambda=0.835;
```

This parameter may be modified by the user. For example, to assume that the state and local total tax burden amounts to 60 percent of the federal tax liability, change line 20 to:

```
%let lambda=0.6;
```

5.5.3. Partial Privatization of Social Security

Social Security's OASI program currently offers a benefit flow which may be partially taxable, as programmed in the tax model. Several reform proposals introduce individual savings accounts into the Social Security program, much like IRAs. To evaluate the after-tax consequences of such proposals, the tax model must be modified to account for income from such individual savings accounts. The proper modification depends on the proposed taxation regime. Consider the following options.

1. Income from individual savings accounts is treated in the same manner as OASI benefits. Under this regime, add estimated income flows from individual savings accounts to OASI benefits, captured by temporary variable `css`. See lines 80-81 of the `computax` macro.

2. Income from individual savings accounts is treated in the same manner as DC pension income. MINT captures income from DC pension accounts through income from assets, part or all of which may be taxable. However, the fraction of income from assets which stems from DC pension accounts is entirely taxable. To treat income from individual savings accounts in the same manner as DC pension income, add estimated income flows from individual savings accounts to temporary variable `cra`, i.e., do not multiply the income by parameter `gamma`. See line 77 of the `computax` macro.

3. Income from individual savings accounts is treated in the same manner as DB pension income. Under this regime, add estimated income flows from individual savings accounts to DB pension benefits, captured by temporary variable `cdb`. See lines 73-74 of the `computax` macro. This treatment is equivalent to treatment like DC pension income.

4. Income from individual savings accounts is exempt from federal income taxation. Under this regime, omit income from individual savings accounts from the tax model.

References

Alho, Juha and Bruce Spencer (1990). "Error Models for Official Mortality Forecasts." Journal of the American Statistical Association, 85:411, 609-616.

Anderson, Joseph (1990). "Micro-Macro Linkages in Economic Models." In Microsimulation Techniques for Tax and Transfer Analysis (G. Lewis and M. Richard, eds.). Washington: The Urban Institute Press.

Bell, Felicitie (1997). "Social Security Area Population Projections: 1997." Social Security Administration Publication Actuarial Study No. 112 (Washington, D.C.). Also at http://www.ssa.gov/OACT/NOTES/AS112/as112cov.html.

Board of Trustees (1997). 1997 Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Disability Insurance Trust Funds. Washington: U.S. Government Printing Office.

Brayton, F. and P. Tinsley, eds. (1996). "A Guide to FRB/US: A Macroeconomic Model of the United States." Finance and Economics Discussion Series, 1996-42. Washington: Federal Reserve Board.

Brayton, F., A. Levin, R. Tryon, and J. Wiliams (forthcoming). "The Evolution of Macro Models at the Federal Reserve Board." Carnegie-Rochester Conference Series on Public Policy.

Brayton, F., E. Mauskopf, E. Reifschneider, D. Tinsley, and J. Williams (1997). "The Role of Expectations in the FRB/US Macroeconomic Model." Federal Reserve Bulletin, April.

Bureau of the Census (1996a). "65+ in the United States." Current Population Reports, P23-190, U.S. Government Printing Office, Washington, DC, 1996.

Bureau of the Census (1996b). "Population Projections of the United States by Age, Sex, Race, and Hispanic Origin: 1995 to 2050." Current Population Reports, P25-1130, U.S. Government Printing Office, Washington, DC, 1996.

Bureau of the Census. 1997. Statistical Abstract of the United States. Government Printing Office, Washington, DC.

Cohen, Lee. 1998. "Stock Portfolio Risks and Returns for Simulation Models." Social Security Administration, Office of Research, Evaluation, and Statistics. Working paper, July 1998.

Congressional Budget Office (1998). "Description of Economic Models." CBO Papers Series. Washington DC: Congressional Budget Office.

Employee Benefit Research Institute/Mathew Greenwald and Associates. 1995. Retirement Confidence Survey. Washington, DC: Employee Benefit Research Institute.

Fair, Ray C. (1994). Testing Macroeconometric Models. Cambridge: Harvard University Press.

General Accounting Office. 1996. 401(k) Pension Plans: Many Take Advantage of Opportunity to Ensure Adequate Retirement Income. HEHS-96-176. August 1996. Government Printing Office, Washington, DC.

Guralnik, Jack, Machiko Yanagishita, and Edward Schneider (1988). "Projecting the Older Population of the United States: Lessons from the Past and Prospects for the Future." The Milbank Quarterly, 66:2, 283-308.

Hubbard, R. G., J. S. Skinner and S. P. Zeldes. 1997. "The Importance of Precautionary Motives for Explaining Individual and Aggregate Saving," in Carnegie-Rochester Conference Series on Public Policy, ed. By A. H. Meltzer and C. I. Plosser, vol. 40, pp. 59-126.

Hurd, Michael J., Lee A. Lillard and Constantijn W. A. Panis. 1998. "An Analysis of the Choice to Cash out, Maintain, or Annuitize Pension Rights Upon Job Change or Retirement." RAND mimeo DRU-1979-DOL.

Ibbotson Associates. 1998. Stocks, Bonds, Bills, and Inflation 1998 Yearbook. Ibbotson Associates, Chicago, Illinois.

Kalbfleisch, John D. and Ross L. Prentice. 1980. The Statistical Analysis of Failure Time Data. John Wiley and Sons, New York.

Keyfitz, Nathan (1981). "The Limits of Population Forecasting." Population and Development Review, 7:4, 579-593.

Klerman, Jacob A. and Constantijn Panis. 1999. *User's Guide to SSA's MINT Tax Model.* Unpublished RAND report, August 1999.

Korenman, Sanders, Noreen Goldman, and Haishan Fu. 1997. "Misclassification Bias in Estimates of Bereavement Effects." American Journal of Epidemiology 145(11): 995-1002.

Land, Kenneth (1986). "Methods for National Population Forecasts: A Review." Journal of the American Statistical Association, 81:396, 888-901.

Lee, Ronald and Lawrence Carter (1992). "Modeling and Forecasting U.S. Mortality." Journal of the American Statistical Association, 87:419, 659-671.

Lee, Ronald and Shripad Tuljapurkar (1994). "Stochastic Population forecasts for the U.S.: Beyond High, Medium and Low." Journal of the American Statistical Association, 87:419, 659-671

Leeper, Eric and Christopher Sims (1994). "Towards a Macroeconometric Model Usable for Policy Analysis." NBER Macroeconomics Annual.

Lewis, Gordon and Richard Michel, eds. (1990). Microsimulation Techniques for Tax and Transfer Analysis. Washington: The Urban Institute Press.

Lillard, Lee A., and Lynn Karoly. 1997. "Marital Status, Earnings and Wealth Accumulation Over the Life Cycle," paper presented to conference on Income, Expenditures & Wealth: Panel Measurements & Behavior, Santa Fe, NM

Lillard, Lee A., and Yoram Weiss. 1997. "Uncertain Health and Survival: Effects on End of Life Consumption." Special Issue: Dynamic Structural Models, Journal of Business and Economic Statistics, Vol. 15(2): 254-268.

Lillard, Lee, and Constantijn Panis. 1996. "Marital Status and Mortality: The Role of Health." Demography, 33(3): 313-327.

Lillard, Lee, and Constantijn Panis. 1998a. "Panel Attrition from the Panel Study of Income Dynamics: Marital Status, Income, and Mortality." Journal of Human Resources 33(2): 437-457.

Lillard, Lee, and Constantijn Panis. 1998b. "Income and Mortality: The Role of Health." Under review.

Lillard, Lee. 1993. "Simultaneous Equations for Hazards: Marriage Duration and Fertility Timing." Journal of Econometrics 56: 189-217.

McCarthy, Kevin F. and George Vernez. 1997. "Immigration in a Changing Economy: California's Experience." Santa Monica, California: RAND MR-154.

Meyer, Laurence (1994). "Comment on 'Towards a Macroeconometric Model Usable for Policy Analysis.'" NBER Macroeconomic Annual.

Monthly Vital Statistics Report Volume 43, No. 12, Supplement (July 14, 1995: "Advance Report of Final Marriage Statistics, 1989 and 1990")

National Center for Health Statistics. 1997. *U.S. Decennial Life Tables for 1989-91. Volume I, Number I, United States Life Tables..* Hyattsville, Maryland.

National Center for Health Statistics. 1998. *Vital statistics of the United States, 1994,* preprint of vol II, mortality, part A sec 6 life tables. Hyattsville, Maryland.

National Institute on Aging (1984). Macroeconomic-Demographic Model. Washington: U.S. Department of Health and Human Services.

O'Shaughnessy, James. 1998. What Works on Wall Street. McGraw-Hill, New York.

OASDI Board of Trustees. 1997. 1997 Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Disability Insurance Trust Funds. Government Printing Office. Washington, DC.

Panis, Constantijn and Lee Lillard. 1996. "Socioeconomic Differentials in the Returns to Social Security." RAND mimeo DRU-1327-NIA.

Papke, Leslie E. 1998. "How Are Participants Investing Their Accounts in Participant-Directed Individual Account Pension Plans?" American Economic Review 88(2); 212-216.

Poterba, J. M., S. F. Venti, and D. A. Wise. 1996. "Personal Retirement Saving Programs and Asset Accumulation: Reconciling the Evidence." NBER Working Paper No. 5599, May 1996.

Poterba, J.M., S. F. Venti, and D. A. Wise. 1995. "Do 401(k) Contributions Crowd Out Other Personal Saving?" Journal of Public Economics 58: 1-32.

Sargent, Robert (1985). "An Expository on Verification and Validation of Simulation Models." In Proceedings of the 1985 Winter Simulation Conference (D. Grants, G. Blais, and S. Solomon, eds.).

Scott, Jason and John Shoven. 1996. "Lump Sum Distributions: Fulfilling the Portability Promise or Eroding Retirement Security?", EBRI Issue Brief, Washington DC: Employee Benefit Research Institute, October.

Smith, James P. 1997. "Marriage, Assets and Savings," paper presented at RAND conference on Reshaping the Family: Social and Economic Changes and Public Policy, Santa Monica, CA.

Social Security Administration. 1996. Annual Statistical Supplement, 1996, to the Social Security Bulletin. US Government Printing Office, Washington, DC.

Sundén, Annika E. and Brian J. Surette. 1998. "Gender Differences in the Allocation of Assets in Retirement Savings Plans." American Economic Review 88(2): 207-211.

Toder, Eric, Cori Uccello, John O'Hare, Melissa Favreault, Caroline Ratcliffe, and Karen Smith. *Modelling Income in the Near Term—Projections of Retirement Income Through 2020 for the 1931-60 Birth Cohorts.* Draft Final Report, April 1999. The Urban Institute, Washington, DC.

Tuljapurkar, Shripad (1992). "Stochastic Population Forecasts and Their Uses." International Journal of Forecasting, 8, 385-391.

Yakoboski, Paul and Jack VanDerhei. 1996. Worker Investment Decisions: An Analysis of Large 401(k) Plan Data. Issue Brief 176, August 1996. Employee Benefits Research Institute. Washington, DC.

Yakoboski, Paul. 1997. "Large Plan Lump-Sums: Rollovers and Cashouts." EBRI Issue Brief 188. Employee Benefit Research Institute, Washington, DC.

Zedlewski, Sheila (1990). "The Development of the Dynamic Simulation of Income Model (DYNASIM)." In Microsimulation Techniques for Tax and Transfer Analysis (G. Lewis and M. Richard, eds.). Washington, DC: The Urban Institute Press.

A. Demographic Projections Appendices

A.1. Overview

This appendix documents all SAS programs that were used to prepare the SIPP data for analysis and to project demographic histories for the simulation sample.

The simulations are based on the 1990, 1991, 1992, and 1993 SIPP panels. While there are differences between these panels, most variables are defined identically and available in all four waves. Data preparation follows the following sequence:

1. For each SIPP panel, read raw (ascii) data from the SIPP Core file and several Topical Modules into SAS data sets.
2. Read raw Numident data matched to the four SIPP panels into one SAS data set.
3. For each SIPP panel, select variables of interest, merge the Core file with the Topical Modules, clean the data, and construct demographic histories.
4. Merge the resulting four (1990-1993) SIPP data sets, impute missing values, merge in Numident information, and process the data.
5. Estimate demographic transition models.
6. Project demographic transitions from the last interview wave through death.

The remaining sections describe the SAS programs for each step. For underlying statistical models and algorithms, please refer to the text in Chapter 2 of this report.

All programs were delivered to SSA on a 100MB ZIP diskette.[49] The file structure contains a directory `Prog`; all SAS programs documented in this appendix, including corresponding `.log` and `.lst` files, are in that `Prog` directory, unless explicitly noted otherwise. Data sets are all in directory `Data` or its subdirectories `Data\1990`, `Data\1991`, `Data\1992`, or `Data\1993`.

[49] Files `mint2.sas` and `durdisab.mac`, discussed in Appendix A.7, were delivered in August 1999 on a 1.44MB diskette.

A.2. Convert SIPP ASCII Data into SAS Data Files

The projections are based on respondents to the 1990, 1991, 1992, and 1993 SIPP that were born in 1926-1965 and that had a positive value of full panel weight (`pnlwgt`) or responded to all survey waves. These respondents were roughly 33-64 years old as of their last interview wave. Since we need to project demographic transitions through the year 2020 (when the oldest respondents will be 89 years old), we need to estimate demographic transition models that apply to all ages, including the oldest old. We therefore processed SIPP records not only for respondents for which projections are required, but for as many respondents as possible.

We used the core files of all SIPP waves (the "Full Panel Research File"), plus topical modules of Waves 2 and 3. The Marital History module of Wave 2 contains respondents' marital history through the Wave 2 interview date. Core files of Wave 3 through the last wave are used to update the marriage history through the last interview wave. The 1990 and 1991 SIPP had eight interview waves; the 1992, ten waves; and the 1993 panel, nine waves.

RAND's data library exists as part of its Unix-based network. Therefore, the raw SIPP data upon which this task was based was processed initially under Unix, using SAS 6.12 for Sun OS. After reading the raw data and creating SAS transport files, all subsequent processing and analysis was done using SAS 6.12 running under Windows NT 4.0. The Unix-SAS programs may be run virtually unchanged on Windows NT. Only data library (directory) references need to be adjusted.

There are separate programs to read the raw data of the four SIPP panels. Their names indicate the SIPP panel to which they apply. For example, `read-fp90.sas` reads the Full Panel research file of the 1990 SIPP; `read-fp91.sas` reads the 1991 SIPP, et cetera. We denote such names in generic notation by `read-fpYY.sas`, where YY=`90`, `91`, `92`, or `93`. The following programs read raw ASCII data and create SAS data sets.

`read-fpYY.sas` – Reads the 1990-1993 full panel research files, and creates SAS transport files. RAND's full panel research files are stored as multiple separate compressed files; the program processes each component one at a time, uncompressing, reading, and concatenating the results into a final file, which is written as a SAS transport file.

input — raw 1990-1993 full panel research files
output — `fp90.xpt`, `fp91.xpt`, `fp92.xpt`, `fp93.xpt`

`read-tm.sas` – Reads topical modules and create SAS transport files. The program consists of a macro, `readtm`, that will read any raw

topical module. The relevant input statement is included from subdirectory `Prog\Include` The arguments to the macro are:
1) wave
2) year
3) lrecl of the raw data file

input — raw topical module files
output — `tmYYwvW.xpt` (note: `YY`=panel year and `W`=wave)

Supporting files and programs

`make-layout.sas` – Program to read a machine-readable SIPP data dictionary and generate a SAS input statement. These can be included in SAS programs as needed using `%include`

`inYYtmw.inc` – SAS input statements generated by `make-layout.sas` These are included in the program `read-tm.sas`

A.3. Convert Numident ASCII Data into SAS Data Files

The RAND removable harddrive on SSA premises in Washington DC contains four ASCII files with SIPP IDs and month of death for those respondents that died before these files were extracted from SSA' s master Numident file. For the 1990 and 1991 SIPP, Numident information was matched in June 1998; for the 1992 and 1993 panels, records were matched in October 1998.

`numident.sas`– Reads the 1990-1993 ASCII files with Numident information, converts the month of death into a SAS date (assuming the death was on the 15th of the month), and appends the four subsamples into one SAS data set.

input — `dod90.txt`, `dod91.txt`, `dod92.txt`, and `dod93.txt`
output — `numident.sd2`

For test purposes on RAND premises, we randomly generate death dates for 2 percent of SIPP respondents:

`fakenumident.sas`– Generates random death dates before June 1998 for 2 percent of SIPP respondents.

input — `sipp2.sd2`
output — `numident.sd2`

When replicating the data preparation sequence on SSA premises, with access to Numident extract files, there is no need to run `fakenumident.sas`

A.4. Process Four SIPP Panels Separately

The 1990-1993 questionnaires and file lay-outs are very similar. We therefore used only one SAS program to carry out a particular task for all 1990-1993 data. For example, edu.sas processes education-related data for all four SIPP panels. It contains a large macro, which is called four times with slightly different parameters. Figure A.1 shows the program flow for initial processing of the Core and Topical Modules data. There are four such sequences, for 1990-1993 data. Note that all SAS data sets have a name that indicates their SIPP panel. For example, edu.sas produces `edu90.sd2`, `edu91.sd2`, `edu92.sd2`, and `edu93.sd2`. The resulting data sets are `sipp90.sd2, sipp91.sd2, sipp92.sd2`, and `sipp93.sd2`.

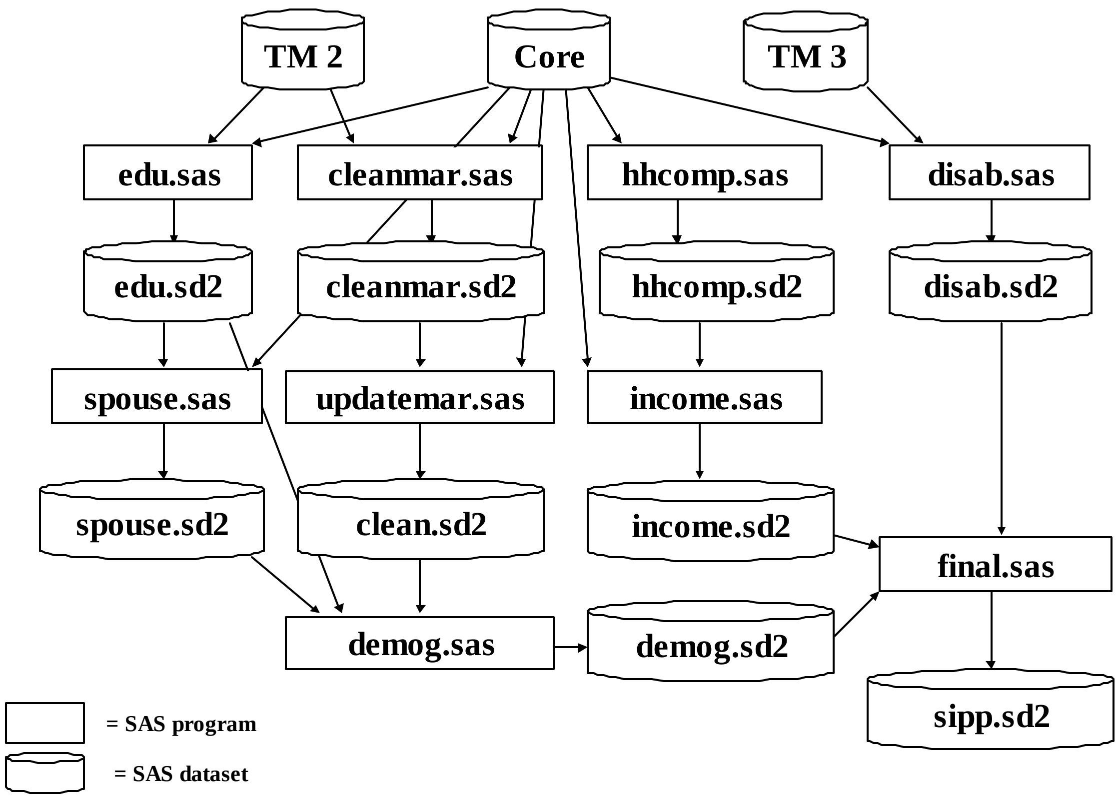

Figure A.1. 1990-1993 SIPP Data Preparation Flow Chart (Part 1)

These programs appear in sequential order, and should be run in the order in which they appear (see flowchart). Note MS-DOS batch program `runall.bat`, which runs all programs in the required sequence. It is located in the mother directory of `Prog` and `Data`.

---

edu.sas - Creates a clean higrade (highest grade completed) variable. The program handles the following problem situations:

1) highest grade drops in successive years.
   - if difference is 1, ignore the issue and take the higher value
   - if the successive grade is coded as 01 or 02, and the former grade is higher, assume data entry error and take the higher value.
   - if the former grade was not completed, assume the respondent overstated and take the previous value.
   - in all other cases, give the respondent the benefit of the doubt.

2) Adjust the higrade variable constructed using higrad1-higrad8 and grdcmp1-grdcmp8 according to Wave 2 topical module variables tm8408 (has ... high school diploma?) and tm8422 (what is ... highest post-hs degree?).

inputs — fpYY.sd2 and tmYYwv2.sd2
output — eduYY.sd2

---

cleanmar.sas - Calculates marriage start and end dates for the first, second, and most recent marriage and howend variables indicating how/whether each marriage ended. It checks to make sure that all marriages end after they start, and that all subsequent marriages begin after a prior marriage ends. The steps are:

1) assigns best guess, start and end of marriage window for each of the marriages described in the wave 2 topical module. This is handled by the macro %make_mar.

2) For cases where the number of marriages is >3, the program imputes marriages 3 to n-1 (the last marriage in the topical module is always the most recent marriage).

3) Next, the program traps problem cases with multiple marriage events in same month; see below.

inputs — fpYY.sd2 and tmYYwv2.sd2
output — cleanYY.sd2

---

hhcomp.sas– Creates a household-level file of household compositions (number of adults and number of children) in each of the 32 panel months. These compositions are merged to the income files in the next step. Normally, the household compositions retained are for the first month in each on the time periods for which incomes are calculated – months 1, 11, and 23, respectively. However, a subset of the population is not present in the sample in those months. Thus, they do not have a household id in these months; therefore, we cannot determine their household compositions. For this group, we define household composition as it existed in the month in which the individual appeared in the sample. This is why it is necessary to build a file of household compositions for all 32 months of the survey.

input — `fpYY.sd2`
output — `hhcompYY.sd2`

---

disab.sas– Builds a file with some indicators of work disability.

inputs — `fpYY.sd2`, `tmYYwv2.sd2`, and `tmYYwv3.sd2`
output — `disabYY.sd2`

---

updatemar.sas– Updates marriage histories based on the `ms*` series. This program starts with the file `cleanmar.sd2`, which contains marriage histories as described by the wave 2 topical module, and updates these marriage histories based on the observable period after the wave 2 topical module (e.g., `ms9` to `ms32`).

Two basic problem situations are handled by the code:

1) The final marriage status based on the wave 2 marriage histories is inconsistent with the full panel variable `ms9`, the month when the topical module was administered. To project marriage histories beyond the wave 23 topical module, these inconsistencies must be reconciled.

   A substantial subset of these cases appear to reflect situations where the respondent legitimately switched status in the month before or after a wave seam, but reported this switch one month off in the full panel file. This tendency to respond to monthly questions retrospectively in blocks of four is known as "seam bias."

The remaining cases followed the basic rule basic rule that the marital status as of the last marriage described by the topical module was correct. The `ms*` series was adjusted accordingly starting in month9 to be consistent with the last observed marital status on the topical module. Processing moved forward from `ms9` to `ms32`, looking for any new changes in marital status. The details of this consistency adjustment are extensively documented in the source code itself.

2) Respondents who did not answer the wave 2 topical module, or people were not-in-universe for the topical module but appear in months 8 and 9 on the full panel file with a marital status. Both these groups have marital histories created based on `ms1-ms32`

After updating the marital histories, the program performs extensive checking to ensure consistency within and across histories.

inputs — `fpYY.sd2`, `tmYYwv2.sd2`, and `cleanYY.sd2`
output — `updateYY.sd2`

`income.sas`- Attaches household, family, and personal incomes to each person' s record. This is done for three time periods – month 1-10, 11-22, and 23-32. Each income is adjusted for the number of months the respondent was actually present in the survey to generate three annual incomes in each of the three periods. Household composition (numbers of adults and children) as of the first month in each of the three time periods is attached from the hhcomp file. Note the comment above that for individuals not present in months 1, 11, and 23, household composition is defined for the month in that time period in which the individual was first observed.

inputs — `fpYY.sd2` and `hhcompYY.sd2`
output — `incomeYY.sd2`

`spouse.sas`- Creates a file with spousal characteristics, such as age, race, ethnicity, and highest grade completed. It also retains complete marriage histories for all spouses, and determines in to which of the respondent' s marriages the spouse belongs.

inputs — `fpYY.sd2`, `eduYY.sd2`, `cleanYY.sd2`, and `disabYY.sd2` output — `spouseYY.sd2`

`demog.sas`- Merges together the core, education, marital history, and spouse file to produce a basic demographic file.

inputs — `fpYY.sd2`, `tmYYwv2.sd2`, `updateYY.sd2`, `eduYY.sd2`, and `spouseYY.sd2`
output — `demogYY.sd2`

`final.sas`- Merges together the demographic file, the incomes file, and the disability file to produce the final analysis file.

inputs — `demogYY.sd2`, `incomeYY.sd2`, and `disabYY.sd2`
output — `sippYY.sd2`

All of the PC-SAS programs make extensive use of a common macro and include library named `common.inc`. This library is `%included` at the start of all the SAS programs. Some of the more commonly used macros in this library are:

`%setup` Parses a sysparm passed in on the command line for the panel year and uses this to decide which year of the SIPP to process. Most useful for batch automation of file construction.

`%lastday`- Array holding the last day in each calendar month

`%datesYY`- Arrays holding the year and month of each of the 32 months in the 1990 and 1991 panels. These vary by rotation group (`moYY` and `yrYY`).

`%setmar`- Macro to assign a marriage event, specifically, the window start, end, and the best guess date of the marriage.

`%checkmar`- Extensive marriage to check the consistency of all marriage events in the marital history file. Checks include

- Check to see that all marriages have a valid start date, end date, and howend variable.
- Check to see if any multiple marriage cases have subsequent marriages starting prior to the end of previous marriages
- Check for consistency in the upper and lower bounds of marriage windows with marriages. This ensures, among other things, that the lower bound starts before the best

guess, and that the best guess occurs before the upper bound.

- Check for consistency of the bounds across successive marriage events to ensure the lower bound of a marriage event window does not start prior to the best guess date of the previous event, and that the upper bound of a marriage event does not occur after the best guess date of the following event.

`%cnt_cris` - Identifies cases where multiple marriage events occur in the same month. Because the default best guess data of a marriage event occurs on the 15th of the month, multiple events would fall on the same day for this subset of cases.

`%fix_cris`- This macro ensures that events in the same month have a time interval between them.

A.5. Append SIPP Panels, Merge Numident Extracts, and Process the Data.

The above sequence of programs generates four data sets, for the 1990, 1991, 1992, and 1993 SIPP (`sipp90.sd2`, `sipp91.sd2`, `sipp92.sd2`, and `sipp93.sd2`, respectively). From this point on, the data sets are appended and they are further processed as one data file. Figure A.2 shows the subsequent programs and flow.

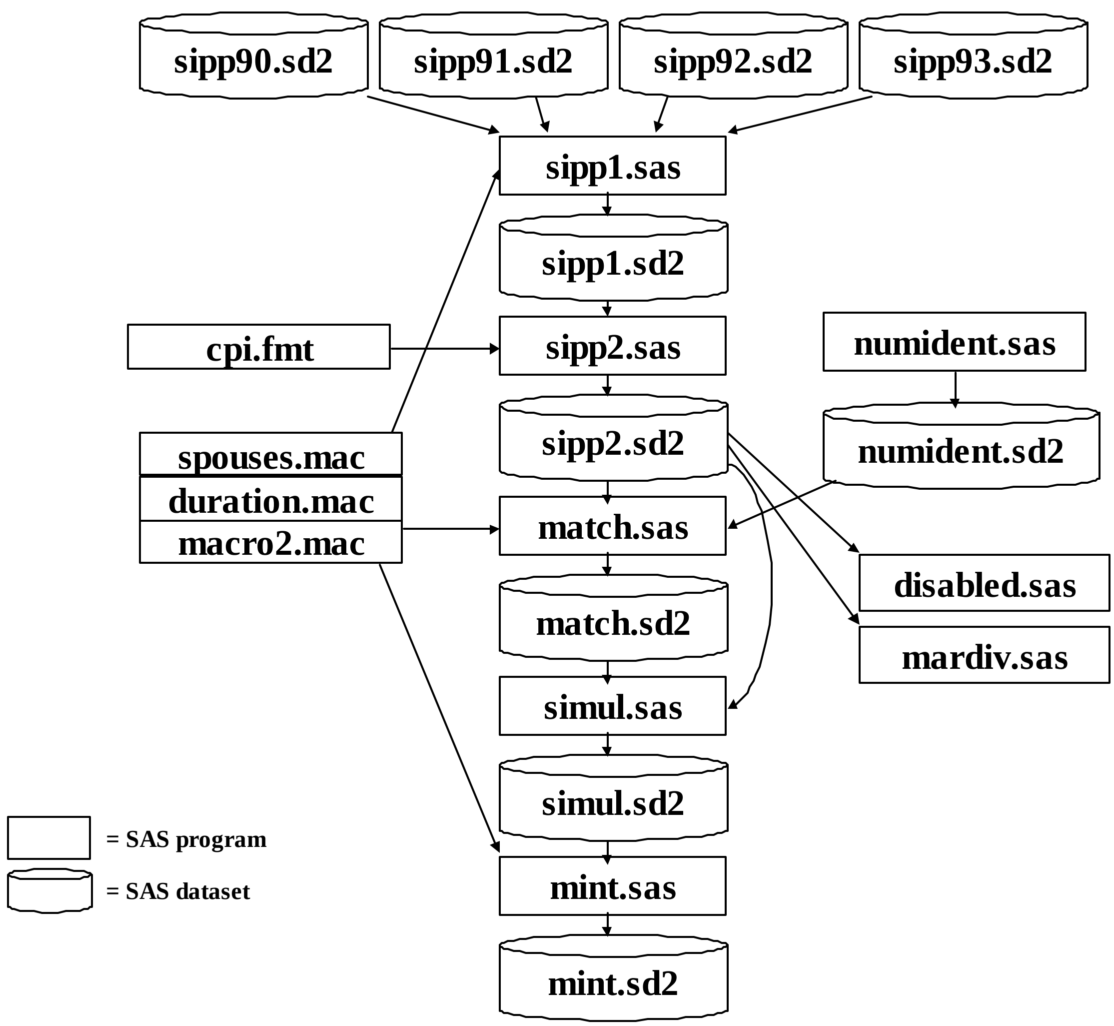

Figure A.2. **1990-93 SIPP Data Preparation Flow Chart (Part 2)**

`sipp1.sas` Appends 1990, 1991, 1992, and 1993 data. Imputes dates of disability onset for respondents, if missing. Also imputes spousal characteristics (birth date, race, Hispanicity, education, disability status, date of disability onset) for all spouses, including those who were never part of the SIPP panels.

Missing values for disability status are imputed by projecting a date of disability onset and comparing it to the respondent's age at the time of the last interview. (Projection of disability onset is done in the same way as mortality, marriage, and divorce dates are projected; see below. These projections require SAS macros in included files `duration.mac` and `macro2.mac`). Imputed spousal age is based on the empirical distribution of age differences between spouses; spousal race, on the bivariate empirical distribution of husbands' and wives' race combinations; spousal Hispanicity, on husbands' and wives' Hispanicity combinations; educational attainment, on spousal education combinations; and spousal disability, on spousal age and other characteristics that predict disability status. Empirical distributions of spousal characteristics were determined by Stata program spouse.do. Its results are incorporated in SAS macro file `spouse.mac`

inputs — `sipp90.sd2`, `sipp91.sd2`, `sipp92.sd2`, `sipp93.sd2`
output — `sipp1.sd2`

`sipp2.sas` Estimates a simple lifecycle model of household log-income based on three annual income measures, and computes the average deviation from the lifecycle for all respondents. This average deviation is taken to be one's permanent income, i.e., one's "long-run" relative deviation from the average income lifecycle pattern. Also see below. Note that household income is converted into 1990 dollars using the Consumer Price Index (CPI). CPI figures are coded in the form of a `proc format` and included in the code through `cpi.fmt`

inputs — `sipp1.sd2` and `cpi.fmt`
output — `sipp2.sd2`

`match.sas` This program merges Numident data with SIPP data. It cleans Numident data: There are duplicate Numident records, and Numident deaths that took place long before the survey. (See Section 2.7.1.) It then projects dates of death (and other demographic transition dates) for the simulation sample. It only projects until June 1998, when the Numident file was created for the 1990 and 1991 SIPP panels, and until October 1998 for the 1992 and 1993 panels. It compares the death rate as of 6/1/1998 or 10/1/1998 to the death rate found in Numident data. It turns out that the Numident records fail to capture all deaths, i.e., the Numident file is based on an incomplete match with the US population. `Match.sas` figures out what fraction of the SIPP simulation sample is not matched (23 percent), and randomly assigns individuals to be matched and not matched. (All individuals for which the Numident file contains a date of death are matched; many but not all of those not in the Numident file are also matched and truly alive; among non-matched individuals there are both living and deceased persons as of 6/1/1998 or 10/1/1998.) The Numident correction is further described in the text.

inputs — `sipp2.sd2`, `numident.sd2`, and SAS macro files `spouse.mac`, `duration.mac`, and `macro2.mac`
output — `match.sd2`

`simul.sas` This program looks up spousal characteristics for those who have spouses in the sample. It then selects the respondents for whom we need to project demographic transition dates, i.e., individuals born in 1926-1965 with positive full panel weight (`pnlwgt`) or who responded to all survey waves.

inputs — `sipp2.sd2` and `match.sd2`
output — `simul.sd2`

This resulting data set, `simul.sd2`, is the basis for demographic transition projections.

A.6. Estimate Demographic Transition Models

Data set `sipp2.sd2` contains demographic histories for all SIPP respondents, including those outside the simulation cohorts. It forms the basis of estimation of demographic models of marriage formation, marriage dissolution, and onset of disability.

`mardiv.sas` This program converts marriage histories into marriage and divorce spells and writes them out in ASCII format.

inputs — `sipp2.sd2`
outputs — `getmar.raw` and `getdiv.raw` (not shown in Figure A.2)

`disab.sas` This program converts disability onset dates into disability spells and writes them out in ASCII format.

inputs — `sipp2.sd2`
output — `disab.raw` (not shown in Figure A.2)

Model estimation is performed outside SAS in aML software. SAS does not support estimation of hazard models of the complexity and richness required by MINT. aML was developed by Lillard and Panis and is commercially available. The contractor PCs on SSA's premises contain copies of aML executable files, `aml.exe` and `raw2aml.exe` The relevant estimation program files are:

`mardiv.r2a` This program reads ASCII data on marriage and divorce spells and converts them into aML data format.

inputs — `getmar.raw` and `getdiv.raw`
output — `mardiv.dat`

`disab.r2a` This program reads ASCII data on disabilityspells and converts them into aML data format.

inputs — `disab.raw`
output — `disab.dat`

`getmar.aml`	This program estimates the male and female models of marriage formation.
inputs —	`mardiv.dat`
output —	`getmar.out`

`getdiv.aml`	This program estimates the male and female models of marriage dissolution.
inputs —	`mardiv.dat`
output —	`getdiv.out`

`disab.aml`	This program estimates the model of onset of disability.
inputs —	`disab.dat`
output —	disab`.out`

Results of estimation are reported in Chapter 2 of this report (Table 2.5, Table 2.7, and Table 2.8). They have been converted into SAS macros `%getmar`, `%getdiv`, and `%disab`, which are contained in included file `duration.mac`

A.7. Project Demographic Transitions

As shown in Figure A.2, the final program in the entire sequence is `mint.sas`

`mint.sas`	Projects dates of marriage and remarriage, divorce, onset of disability, and death. Special care is taken to ensure that the dates of death of individuals appearing in Numident files are projected correctly, and that the overall death rate corresponds to Vital Statistics. Special care is also taken to ensure spousal consistency of divorce, widowhood, and death dates. The output file, `mint.sd2`, is further described below. In addition, `mint.sas` generates `urban.sd2`. It is identical to `mint.sd2`, but also includes projections for individuals born in 1926-1929 and 1961-1965. It is for use by The Urban Institute/Brookings Institution only.
inputs —	`simul.sd2` and SAS macro files `spouse.mac`, `duration.mac`, and `macro2.mac`
output —	`mint.sd2` and `urban.sd2`. (The latter data set is not shown in Figure A.2.)

Mortality as a Function of Disability Status

By default, the projection algorithm in `mint.sas` does not take account of disability status in longevity projections, but is easily modified to do so. Program `mint2.sas` is based on a mortality specification which does control for disability status. It is an alternative to `mint.sas`.[50]

The estimates of the mortality specification with control for disability status are shown in Table 2.13 (page 17); they are corrected for differences between the PSID and Vital Statistics, as explained in Section 2.2.3. Program `mint.sas` includes file `duration.mac` with macros for drawing longevity durations; similarly, `mint2.sas` includes the longevity macros of `durdisab.mac`.

`mint2.sas`	Projects dates of marriage and remarriage, divorce, onset of disability, and death. It differs from `mint.sas` in that it

[50] This projection program is not part of the official set of programs. It is not included on the 100MB ZIP disk on which all programs were transferred from RAND to SSA, but was delivered separately on a 1.44MB diskette in August 1999.

	includes the longevity macros embedded in file `durdisab.mac` (instead of `duration.mac`). These longevity macros account for disability status. It generates `mint2.sd2` (1931-1960 cohorts) and `urban2.sd2` (1926-1965 cohorts).
inputs —	`simul.sd2` and SAS macro files `spouse.mac`, `durdisab.mac`, and `macro2.mac`
output —	`mint2.sd2` and `urban2.sd2`

A.8. Simulation Data Set

SAS data set `mint.sd2` contains projections of individual demographic transitions. Each SAS observation corresponds to one individual. We now describe the key simulation outcome variables.

The main variables of interest are `sipp`, `id`, and `surv2020`, `stat62`, `stat67`, and `stat2020`.

`sipp`	1990=from 1990 SIPP; 1991=from 1991 SIPP; 1992=from 1992 SIPP; 1993=from 1993 SIPP
`id`	100000*`ppid`+1000*`ppent`+`ppnum` The respondent ID, in numerical format
`surv2020`	Probability of surviving until 1/1/2020
`stat62`	Demographic status at the respondent's 62nd birthday.
`stat67`	Demographic status at the respondent's 67th birthday.
`stat2020`	Demographic status at January 1, 2020 0 = never married 1 = married 2 = widowed 3 = divorced 4 = deceased

Variable `deathdte` contains the projected date of death. For all respondents, marriage transitions are simulated until death, so there is a nonmissing death date for every record.

`deathdte`	Projected date of death

This and all other date variables are in the standard SAS date format, i.e., internally they represent the number of days since 1/1/1960. They are all formatted with `date7.` or `date9.` formats, so that the human eye can interpret them readily.

Other variables of interest are:

`disabled`	Functionally disabled 0 if still in good health or died while still in good health The definition of disability is a self-reported health or mental condition which limits the amount or kind of work that the respondent can do.

For projection purposes, we assume that it is an absorbing state, i.e., one never moves from disabled to healthy.

`disabdte` Date became disabled
This variable may be missing if the person reported being disabled in the topical module (Wave 3), but did not give a date of onset. It is also missing if the person did not fall into disability (disabled=0).

`nummar` number of marriages
To be precise, number of weddings. Takes values 0-8.

`marb_1-marb_12` Wedding dates (`marb_9-marb_12` are always missing)
`mare_1-mare_12` Marriage end dates (`mare_9-mare_12` are always missing)
`howend1-howend12` Disposition of marriage:
0 = still married as of the last date (happens since all eventually die)
1 = divorce
2 = widowhood
3 = own death

`educ` Education:
1 = high school drop-out
2 = high school graduate
3 = college graduate

`hisp` Hispanic (14<=ethnicty<=20)

`male` Male

`perminc` Permanent income
Measured as the person-specific deviation from a life-cycle pattern in log-household income, controlling only for age, sex, marital status, and household composition.

`race` Race:
1 = White
2 = Black
3 = Native American
4 = Asian

For all spouses that were present in the survey, similar demographic information is available. In addition, we imputed such information for all spouses that married after the last survey wave. Spousal characteristics are in:

`sprace1-sprace12` Spousal race
`sphisp1-sphisp12` Spousal hispanicity
`speduc1-speduc12` Spousal education

`spinc1-spinc12`	Spousal permanent income
`spdisab1-spdisab12`	Spousal disability
`spdisdt1-spdisdt12`	Date of onset of spousal disability

Note that the projections are carried out until all respondents were deceased, i.e., well beyond the year 2020. The current projections enable a quick look-up of demographic status as of any date. A particularly useful routine for this purpose is in SAS macro `%figstat`:

```
/* Define macro to figure out demographic status as of date */
%macro figstat(date,status);
   /*   0 = never married */
   /*   1 = married       */
   /*   2 = widowed       */
   /*   3 = divorced      */
   /*   4 = deceased      */
   &status=.;
   if (deathdte^=. and &date>=deathdte) then &status=4;  /* deceased */
   else if (nummar=0) then &status=0; /* never married */
   else if (&date<marb(1)) then &status=0;  /* never married */
   else do;
      do ii=1 to nummar while (&status=.);
         if (marb(ii)<=&date<mare(ii)) then do;
            &status=1;  /* married */
         end; else if (&date<marb(ii)) then do;
            if (howend(ii-1)=0) then &status=1;      /* married */
            else if (howend(ii-1)=1) then &status=3; /* divorced */
            else if (howend(ii-1)=2) then &status=2; /* widowed */
            else if (howend(ii-1)=3) then &status=4; /* deceased */
            else put “Error 1 in program logic!”;
         end; else if (ii=nummar and &date>=mare(nummar)) then do;
            if (howend(ii)=0) then &status=1;      /* married */
            else if (howend(ii)=1) then &status=3; /* divorced */
            else if (howend(ii)=2) then &status=2; /* widowed */
            else if (howend(ii)=3) then &status=4; /* deceased */
            else put “Error 2 in program logic!”;
         end;
      end;
   end;
   if (&status=.) then put “Error 3 in program logic!” _all_;
%mend;
```

This routine is also contained in `mint.sas`. Its use is as follows. Suppose one wants a respondent’ s demographic status as of 1/1/2005. The code is:

```
date = mdy(1,1,2005);
%figstat(date,stat2005);
```

The demographic status will now be stored in variable `stat2005`. Its format is the same as that of `stat62`, `stat67`, and `stat2020`; it is often convenient to add the statement “`format stat2005 demostat.;`”, so that `stat2005`'s values are easily interpreted by the human eye.

We removed Numident information from the resulting SAS data set with projections, `mint.sd2` It does therefore not contain any confidential information.

Note that both `mint.sd2`and `urban.sd2`contain imputed spousal characteristics for all spouses, including those who were married to SIPP respondents before the SIPP surveys.

Finally, we present the output of a `proc contents`and a `proc means`on the simulation data set, `mint.sd2`

```
CONTENTS PROCEDURE

Data Set Name: DDOUT.MINT                              Observations:         84497
Member Type:   DATA                                    Variables:            138
Engine:        V612                                    Indexes:              0
Created:       13:44 Thursday, March 4, 1999           Observation Length:   600
Last Modified: 13:51 Thursday, March 4, 1999           Deleted Observations: 0
Protection:                                            Compressed:           NO
Data Set Type:                                         Sorted:               NO
Label:

-----Engine/Host Dependent Information-----

Data Set Page Size:       16384
Number of Data Set Pages: 3131
File Format:              607
First Data Page:          2
Max Obs per Page:         27
Obs in First Data Page:   25

                   -----Alphabetic List of Variables and Attributes-----

  #   Variable    Type   Len   Pos   Format      Label
-------------------------------------------------------------------------------------------
 29   BRTHDATE    Num      4   120   DATE9.      best guess birthdate
132   DEATHDTE    Num      4   572   DATE9.      Projected date of death
 72   DISABDTE    Num      4   324   DATE9.      Date became disabled
 71   DISABLED    Num      4   320               Functionally disabled
 92   EDUC        Num      4   408               Education (dropout-graduate-college)
 93   HISP        Num      4   412               Hispanic
133   HORIZON     Num      4   576   DATE9.      Projection horizon
  3   HOWEND1     Num      4    16   HOWEND.     how marriage 1 ended
  4   HOWEND2     Num      4    20   HOWEND.     how marriage 2 ended
  5   HOWEND3     Num      4    24   HOWEND.     how marriage 3 ended
  6   HOWEND4     Num      4    28   HOWEND.     how marriage 4 ended
  7   HOWEND5     Num      4    32   HOWEND.     how marriage 5 ended
  8   HOWEND6     Num      4    36   HOWEND.     how marriage 6 ended
  9   HOWEND7     Num      4    40   HOWEND.     how marriage 7 ended
 10   HOWEND8     Num      4    44   HOWEND.     how marriage 8 ended
 96   HOWEND9     Num      4   428
 97   HOWEND10    Num      4   432
 98   HOWEND11    Num      4   436
 99   HOWEND12    Num      4   440
  2   ID          Num      8     8   14.         ID=100000*ppid+1000*ppent+ppnum
 73   LASTINT     Num      8   328   DATE7.      Last interview date
 91   MALE        Num      4   404               Male
 11   MARB_1      Num      4    48   DATE9.      best guess date of marriage 1
 12   MARB_2      Num      4    52   DATE9.      best guess date of marriage 2
 13   MARB_3      Num      4    56   DATE9.      best guess date of marriage 3
 14   MARB_4      Num      4    60   DATE9.      best guess date of marriage 4
 15   MARB_5      Num      4    64   DATE9.      best guess date of marriage 5
 16   MARB_6      Num      4    68   DATE9.      best guess date of marriage 6
 17   MARB_7      Num      4    72   DATE9.      best guess date of marriage 7
 18   MARB_8      Num      4    76   DATE9.      best guess date of marriage 8
100   MARB_9      Num      4   444   DATE9.
101   MARB_10     Num      4   448   DATE9.
102   MARB_11     Num      4   452   DATE9.
103   MARB_12     Num      4   456   DATE9.
 19   MARE_1      Num      4    80   DATE9.      best guess date of end of marriage 1
 20   MARE_2      Num      4    84   DATE9.      best guess date of end of marriage 2
 21   MARE_3      Num      4    88   DATE9.      best guess date of end of marriage 3
 22   MARE_4      Num      4    92   DATE9.      best guess date of end of marriage 4
 23   MARE_5      Num      4    96   DATE9.      best guess date of end of marriage 5
 24   MARE_6      Num      4   100   DATE9.      best guess date of end of marriage 6
 25   MARE_7      Num      4   104   DATE9.      best guess date of end of marriage 7
 26   MARE_8      Num      4   108   DATE9.      best guess date of end of marriage 8
104   MARE_9      Num      4   460   DATE9.
105   MARE_10     Num      4   464   DATE9.
```

#	Variable	Type	Len	Pos	Format	Label
106	MARE_11	Num	4	468	DATE9.	
107	MARE_12	Num	4	472	DATE9.	
28	MARQUAL	Num	4	116		marriage history quality
27	NUMMAR	Num	4	112		number of marriages
94	PERMINC	Num	8	416		Permanent income
1	PNLWGT	Num	8	0		Full panel weight
70	RACE	Num	4	316		Race (white-black-native-asian)
90	SIPP	Num	4	400		SIPP wave (1990, 1991, 1992, 1993)
38	SPBDAT1	Num	4	188	DATE9.	marriage 1 - spouse birth date
39	SPBDAT2	Num	4	192	DATE9.	marriage 2 - spouse birth date
40	SPBDAT3	Num	4	196	DATE9.	marriage 3 - spouse birth date
41	SPBDAT4	Num	4	200	DATE9.	marriage 4 - spouse birth date
42	SPBDAT5	Num	4	204	DATE9.	marriage 5 - spouse birth date
43	SPBDAT6	Num	4	208	DATE9.	marriage 6 - spouse birth date
44	SPBDAT7	Num	4	212	DATE9.	marriage 7 - spouse birth date
45	SPBDAT8	Num	4	216	DATE9.	marriage 8 - spouse birth date
108	SPBDAT9	Num	4	476	DATE9.	
109	SPBDAT10	Num	4	480	DATE9.	
110	SPBDAT11	Num	4	484	DATE9.	
111	SPBDAT12	Num	4	488	DATE9.	
46	SPDISA1	Num	4	220		marriage 1 - spousal disability
47	SPDISA2	Num	4	224		marriage 2 - spousal disability
48	SPDISA3	Num	4	228		marriage 3 - spousal disability
49	SPDISA4	Num	4	232		marriage 4 - spousal disability
50	SPDISA5	Num	4	236		marriage 5 - spousal disability
51	SPDISA6	Num	4	240		marriage 6 - spousal disability
52	SPDISA7	Num	4	244		marriage 7 - spousal disability
53	SPDISA8	Num	4	248		marriage 8 - spousal disability
124	SPDISA9	Num	4	540		
125	SPDISA10	Num	4	544		
126	SPDISA11	Num	4	548		
127	SPDISA12	Num	4	552		
54	SPDISD1	Num	4	252	DATE9.	marriage 1 - spousal disab onset date
55	SPDISD2	Num	4	256	DATE9.	marriage 2 - spousal disab onset date
56	SPDISD3	Num	4	260	DATE9.	marriage 3 - spousal disab onset date
57	SPDISD4	Num	4	264	DATE9.	marriage 4 - spousal disab onset date
58	SPDISD5	Num	4	268	DATE9.	marriage 5 - spousal disab onset date
59	SPDISD6	Num	4	272	DATE9.	marriage 6 - spousal disab onset date
60	SPDISD7	Num	4	276	DATE9.	marriage 7 - spousal disab onset date
61	SPDISD8	Num	4	280	DATE9.	marriage 8 - spousal disab onset date
128	SPDISD9	Num	4	556	DATE9.	
129	SPDISD10	Num	4	560	DATE9.	
130	SPDISD11	Num	4	564	DATE9.	
131	SPDISD12	Num	4	568	DATE9.	
82	SPEDUC1	Num	4	368		Education (1-2-3) spouse 1
83	SPEDUC2	Num	4	372		Education (1-2-3) spouse 2
84	SPEDUC3	Num	4	376		Education (1-2-3) spouse 3
85	SPEDUC4	Num	4	380		Education (1-2-3) spouse 4
86	SPEDUC5	Num	4	384		Education (1-2-3) spouse 5
87	SPEDUC6	Num	4	388		Education (1-2-3) spouse 6
88	SPEDUC7	Num	4	392		Education (1-2-3) spouse 7
89	SPEDUC8	Num	4	396		Education (1-2-3) spouse 8
120	SPEDUC9	Num	4	524		
121	SPEDUC10	Num	4	528		
122	SPEDUC11	Num	4	532		
123	SPEDUC12	Num	4	536		
74	SPHISP1	Num	4	336		Hispanicity spouse 1
75	SPHISP2	Num	4	340		Hispanicity spouse 2
76	SPHISP3	Num	4	344		Hispanicity spouse 3
77	SPHISP4	Num	4	348		Hispanicity spouse 4
78	SPHISP5	Num	4	352		Hispanicity spouse 5
79	SPHISP6	Num	4	356		Hispanicity spouse 6
80	SPHISP7	Num	4	360		Hispanicity spouse 7
81	SPHISP8	Num	4	364		Hispanicity spouse 8
116	SPHISP9	Num	4	508		
117	SPHISP10	Num	4	512		

#	Variable	Type	Len	Pos	Format	Label
118	SPHISP11	Num	4	516		
119	SPHISP12	Num	4	520		
30	SPID1	Num	8	124	14.	marriage 1 - spouse id
31	SPID2	Num	8	132	14.	marriage 2 - spouse id
32	SPID3	Num	8	140	14.	marriage 3 - spouse id
33	SPID4	Num	8	148	14.	marriage 4 - spouse id
34	SPID5	Num	8	156	14.	marriage 5 - spouse id
35	SPID6	Num	8	164	14.	marriage 6 - spouse id
36	SPID7	Num	8	172	14.	marriage 7 - spouse id
37	SPID8	Num	8	180	14.	marriage 8 - spouse id
62	SPRACE1	Num	4	284		marriage 1 - spouse race
63	SPRACE2	Num	4	288		marriage 2 - spouse race
64	SPRACE3	Num	4	292		marriage 3 - spouse race
65	SPRACE4	Num	4	296		marriage 4 - spouse race
66	SPRACE5	Num	4	300		marriage 5 - spouse race
67	SPRACE6	Num	4	304		marriage 6 - spouse race
68	SPRACE7	Num	4	308		marriage 7 - spouse race
69	SPRACE8	Num	4	312		marriage 8 - spouse race
112	SPRACE9	Num	4	492		
113	SPRACE10	Num	4	496		
114	SPRACE11	Num	4	500		
115	SPRACE12	Num	4	504		
95	STAT	Num	4	424	DEMOSTAT.	Demographic status at last interview
135	STAT62	Num	4	584	DEMOSTAT.	
136	STAT67	Num	4	588	DEMOSTAT.	
137	STAT2020	Num	4	592	DEMOSTAT.	
138	STATHOR	Num	4	596	DEMOSTAT.	
134	SURV2020	Num	4	580		Probability of surviving until 1/1/2020

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
PNLWGT	Full panel weight	84497	4573.89	3581.45	0	85094.43
ID	ID=100000*ppid+1000*ppent+ppnum	84497	5.1374039E13	2.8647759E13	107403411101	9.9994431E13
HOWEND1	how marriage 1 ended	79610	1.9058410	0.8091367	1.0000000	3.0000000
HOWEND2	how marriage 2 ended	24944	1.9760263	0.8200487	1.0000000	3.0000000
HOWEND3	how marriage 3 ended	6454	1.9067245	0.8264803	1.0000000	3.0000000
HOWEND4	how marriage 4 ended	1868	1.9652034	0.8066204	1.0000000	3.0000000
HOWEND5	how marriage 5 ended	505	1.9485149	0.8221408	1.0000000	3.0000000
HOWEND6	how marriage 6 ended	163	1.8466258	0.8133365	1.0000000	3.0000000
HOWEND7	how marriage 7 ended	53	1.8679245	0.7853891	1.0000000	3.0000000
HOWEND8	how marriage 8 ended	14	1.9285714	0.6157279	1.0000000	3.0000000
MARB_1	best guess date of marriage 1	79610	4871.25	4291.16	-6956.00	37017.00
MARB_2	best guess date of marriage 2	24944	10410.03	5582.69	-4491.00	38958.00
MARB_3	best guess date of marriage 3	6454	13087.24	5705.35	-3791.00	35234.00
MARB_4	best guess date of marriage 4	1868	15073.53	5997.96	-1661.00	36067.00
MARB_5	best guess date of marriage 5	505	15451.91	6358.96	1423.00	33227.00
MARB_6	best guess date of marriage 6	163	14412.88	5999.97	2571.00	29273.00
MARB_7	best guess date of marriage 7	53	14159.42	5944.86	7044.00	29318.00
MARB_8	best guess date of marriage 8	14	17209.79	3985.24	13091.00	27203.00
MARE_1	best guess date of end of marriage 1	79610	16619.04	8684.71	-5099.00	39700.00
MARE_2	best guess date of end of marriage 2	24944	18484.95	7526.16	-3852.00	39160.00
MARE_3	best guess date of end of marriage 3	6454	19017.35	6877.70	-2178.00	36521.00
MARE_4	best guess date of end of marriage 4	1868	20094.98	6913.69	849.0000000	37472.00
MARE_5	best guess date of end of marriage 5	505	19734.85	7301.15	1997.00	36433.00
MARE_6	best guess date of end of marriage 6	163	18925.36	7032.53	6038.00	31938.00
MARE_7	best guess date of end of marriage 7	53	18443.17	6495.89	7805.00	30246.00
MARE_8	best guess date of end of marriage 8	14	22208.43	3636.08	16548.00	28249.00
NUMMAR	number of marriages	84497	1.3445566	0.8014126	0	8.0000000
MARQUAL	marriage history quality	84497	0.6403659	1.7235828	0	7.0000000
BRTHDATE	best guess birthdate	84497	-4212.37	3044.45	-10578.00	349.0000000
SPID1	marriage 1 - spouse id	39743	5.1215554E13	2.8588337E13	107406911102	9.9994431E13
SPID2	marriage 2 - spouse id	10213	5.1198368E13	2.864344E13	112300311101	9.9994429E13
SPID3	marriage 3 - spouse id	1792	5.1233836E13	2.8914051E13	198622211101	9.9994429E13
SPID4	marriage 4 - spouse id	254	4.8339665E13	2.9858601E13	575426311101	9.9194446E13
SPID5	marriage 5 - spouse id	67	6.0460031E13	2.9445477E13	635891111102	9.9494476E13
SPID6	marriage 6 - spouse id	33	5.731489E13	2.880638E13	9.4074854E12	9.9170188E13
SPID7	marriage 7 - spouse id	20	6.7885828E13	2.600428E13	1.2600019E13	9.9594409E13
SPID8	marriage 8 - spouse id	0	.	.	.	.
SPBDAT1	marriage 1 - spouse birth date	79610	-4829.03	3496.93	-18799.00	6985.00
SPBDAT2	marriage 2 - spouse birth date	24944	-4595.34	3507.97	-19040.00	6314.00
SPBDAT3	marriage 3 - spouse birth date	6454	-4705.88	3516.12	-18858.00	4732.00
SPBDAT4	marriage 4 - spouse birth date	1868	-4617.88	3294.82	-18585.00	6740.00
SPBDAT5	marriage 5 - spouse birth date	505	-4798.90	3569.48	-16727.00	3241.00
SPBDAT6	marriage 6 - spouse birth date	163	-5904.06	3812.34	-16940.00	2967.00
SPBDAT7	marriage 7 - spouse birth date	53	-6485.55	4000.27	-16026.00	2296.00
SPBDAT8	marriage 8 - spouse birth date	14	-6688.57	2645.70	-11674.00	-2266.00
SPDISA1	marriage 1 - spousal disability	79610	0.5611230	0.4962530	0	1.0000000
SPDISA2	marriage 2 - spousal disability	24944	0.6218329	0.4849393	0	1.0000000
SPDISA3	marriage 3 - spousal disability	6454	0.6501394	0.4769627	0	1.0000000
SPDISA4	marriage 4 - spousal disability	1868	0.7103854	0.4537049	0	1.0000000
SPDISA5	marriage 5 - spousal disability	505	0.7148515	0.4519328	0	1.0000000
SPDISA6	marriage 6 - spousal disability	163	0.7055215	0.4572126	0	1.0000000
SPDISA7	marriage 7 - spousal disability	53	0.7735849	0.4225158	0	1.0000000
SPDISA8	marriage 8 - spousal disability	14	0.9285714	0.2672612	0	1.0000000
SPDISD1	marriage 1 - spousal disab onset date	44671	15145.79	5894.90	-16163.00	30683.00
SPDISD2	marriage 2 - spousal disab onset date	15511	15041.25	5980.08	-12632.00	30252.00
SPDISD3	marriage 3 - spousal disab onset date	4196	14604.21	6088.44	-14809.00	29361.00
SPDISD4	marriage 4 - spousal disab onset date	1327	15108.18	5870.37	-6391.00	27478.00
SPDISD5	marriage 5 - spousal disab onset date	361	15001.17	6098.07	-4747.00	26612.00
SPDISD6	marriage 6 - spousal disab onset date	115	14037.59	6715.01	-7167.00	25814.00
SPDISD7	marriage 7 - spousal disab onset date	41	12870.85	5628.80	-3695.00	22528.00
SPDISD8	marriage 8 - spousal disab onset date	13	11602.46	6664.79	-3788.00	20831.00
SPRACE1	marriage 1 - spouse race	79610	1.2097350	0.6122256	1.0000000	4.0000000
SPRACE2	marriage 2 - spouse race	24944	1.1588759	0.5153061	1.0000000	4.0000000
SPRACE3	marriage 3 - spouse race	6454	1.1310815	0.4705993	1.0000000	4.0000000
SPRACE4	marriage 4 - spouse race	1868	1.1167024	0.4443051	1.0000000	4.0000000
SPRACE5	marriage 5 - spouse race	505	1.0851485	0.3866282	1.0000000	4.0000000

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
SPRACE6	marriage 6 - spouse race	163	1.0797546	0.4003331	1.0000000	4.0000000
SPRACE7	marriage 7 - spouse race	53	1.0943396	0.4049763	1.0000000	3.0000000
SPRACE8	marriage 8 - spouse race	14	1.0714286	0.2672612	1.0000000	2.0000000
RACE	Race (white-black-native-asian)	84497	1.2196409	0.6144242	1.0000000	4.0000000
DISABLED	Functionally disabled	84497	0.8801496	0.3247884	0	1.0000000
DISABDTE	Date became disabled	74370	17067.21	4830.53	-9915.00	28896.00
LASTINT	Last interview date	84497	12443.78	527.0250184	11003.00	13163.00
SPHISP1	Hispanicity spouse 1	79610	0.0856551	0.2798558	0	1.0000000
SPHISP2	Hispanicity spouse 2	24944	0.0702774	0.2556191	0	1.0000000
SPHISP3	Hispanicity spouse 3	6454	0.0497366	0.2174171	0	1.0000000
SPHISP4	Hispanicity spouse 4	1868	0.0396146	0.1951041	0	1.0000000
SPHISP5	Hispanicity spouse 5	505	0.0257426	0.1585234	0	1.0000000
SPHISP6	Hispanicity spouse 6	163	0.0368098	0.1888749	0	1.0000000
SPHISP7	Hispanicity spouse 7	53	0	0	0	0
SPHISP8	Hispanicity spouse 8	14	0	0	0	0
SPEDUC1	Education (1-2-3) spouse 1	79610	2.1007286	0.6339598	1.0000000	3.0000000
SPEDUC2	Education (1-2-3) spouse 2	24944	2.0657473	0.6147427	1.0000000	3.0000000
SPEDUC3	Education (1-2-3) spouse 3	6454	2.0271150	0.6070834	1.0000000	3.0000000
SPEDUC4	Education (1-2-3) spouse 4	1868	1.9817987	0.6085365	1.0000000	3.0000000
SPEDUC5	Education (1-2-3) spouse 5	505	1.9405941	0.5845528	1.0000000	3.0000000
SPEDUC6	Education (1-2-3) spouse 6	163	1.9631902	0.5867842	1.0000000	3.0000000
SPEDUC7	Education (1-2-3) spouse 7	53	1.9056604	0.6868028	1.0000000	3.0000000
SPEDUC8	Education (1-2-3) spouse 8	14	1.7857143	0.6992932	1.0000000	3.0000000
SIPP	SIPP wave (1990, 1991, 1992, 1993)	84497	1991.48	1.1633771	1990.00	1993.00
MALE	Male	84497	0.4806088	0.4996268	0	1.0000000
EDUC	Education (dropout-graduate-college)	84497	2.1029031	0.6204388	1.0000000	3.0000000
HISP	Hispanic	84497	0.0866540	0.2813290	0	1.0000000
PERMINC	Permanent income	84497	-0.2231042	1.4425458	-11.0047224	2.2976320
STAT	Demographic status at last interview	84497	1.2112856	0.8070219	0	3.0000000
HOWEND9		0	.	.	.	.
HOWEND10		0	.	.	.	.
HOWEND11		0	.	.	.	.
HOWEND12		0	.	.	.	.
MARB_9		0	.	.	.	.
MARB_10		0	.	.	.	.
MARB_11		0	.	.	.	.
MARB_12		0	.	.	.	.
MARE_9		0	.	.	.	.
MARE_10		0	.	.	.	.
MARE_11		0	.	.	.	.
MARE_12		0	.	.	.	.
SPBDAT9		0	.	.	.	.
SPBDAT10		0	.	.	.	.
SPBDAT11		0	.	.	.	.
SPBDAT12		0	.	.	.	.
SPRACE9		0	.	.	.	.
SPRACE10		0	.	.	.	.
SPRACE11		0	.	.	.	.
SPRACE12		0	.	.	.	.
SPHISP9		0	.	.	.	.
SPHISP10		0	.	.	.	.
SPHISP11		0	.	.	.	.
SPHISP12		0	.	.	.	.
SPEDUC9		0	.	.	.	.
SPEDUC10		0	.	.	.	.
SPEDUC11		0	.	.	.	.
SPEDUC12		0	.	.	.	.
SPDISA9		0	.	.	.	.
SPDISA10		0	.	.	.	.
SPDISA11		0	.	.	.	.
SPDISA12		0	.	.	.	.
SPDISD9		0	.	.	.	.
SPDISD10		0	.	.	.	.
SPDISD11		0	.	.	.	.
SPDISD12		0	.	.	.	.
DEATHDTE	Projected date of death	84497	25273.90	6285.16	11003.00	44602.00

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
HORIZON	Projection horizon	84497	51135.00	0	51135.00	51135.00
SURV2020	Probability of surviving until 1/1/2020	84497	0.7077480	0.1981006	1.6576068E-6	0.9671321
STAT62		84497	1.7098477	1.1661334	0	4.0000000
STAT67		84497	1.9229085	1.2624396	0	4.0000000
STAT2020		84497	2.2258897	1.3637265	0	4.0000000
STATHOR		84497	4.0000000	0	4.0000000	4.0000000

B. Tax Model Appendices

B.1. Source Code of `cpi.sas`

This appendix lists the source code of `cpi.sas` This file defines a format to map calendar years into the corresponding (projected) Consumer Price Index for Urban Wage Earners and Clerical Workers (CPI). Future values are based on intermediate assumptions of OASDI Board of Trustees (1998).

Source: Years 1950 through 1997 from Table 3.C4 of the 1998 Annual Statistical Supplement (rescaled such that 1998 is 100). Projections for years 1997 through 2075 from Table III.B1 of OASDI Board of Trustees (1998). Projections for individual years after 2007 were interpolated linearly.

Source Code

```
proc format;
value cpi
  1950 = '15.29'
  1951 = '16.20'
  1952 = '16.32'
  1953 = '16.45'
  1954 = '16.32'
  1955 = '16.39'
  1956 = '16.87'
  1957 = '17.36'
  1958 = '17.67'
  1959 = '17.98'
  1960 = '18.22'
  1961 = '18.34'
  1962 = '18.59'
  1963 = '18.89'
  1964 = '19.08'
  1965 = '19.44'
  1966 = '20.12'
  1967 = '20.73'
  1968 = '21.70'
  1969 = '23.05'
  1970 = '24.33'
  1971 = '25.13'
  1972 = '25.98'
  1973 = '28.25'
  1974 = '31.73'
  1975 = '33.93'
  1976 = '35.58'
  1977 = '37.97'
  1978 = '41.39'
  1979 = '46.89'
  1980 = '52.76'
  1981 = '57.47'
```

```
1982 = '59.67'
1983 = '61.94'
1984 = '64.38'
1985 = '66.83'
1986 = '67.56'
1987 = '70.56'
1988 = '73.67'
1989 = '77.10'
1990 = '81.81'
1991 = '84.31'
1992 = '86.76'
1993 = '89.14'
1994 = '91.53'
1995 = '93.85'
1996 = '96.97'
1997 = '98.62'
1998 = '100.00'
1999 = '102.38'
2000 = '105.01'
2001 = '107.82'
2002 = '110.89'
2003 = '114.33'
2004 = '118.02'
2005 = '122.03'
2006 = '126.28'
2007 = '130.73'
2008 = '135.47'
2009 = '140.20'
2010 = '144.94'
2011 = '150.38'
2012 = '155.82'
2013 = '161.26'
2014 = '166.70'
2015 = '172.14'
2016 = '178.60'
2017 = '185.06'
2018 = '191.53'
2019 = '197.99'
2020 = '204.45'
2021 = '212.12'
2022 = '219.80'
2023 = '227.47'
2024 = '235.15'
2025 = '242.82'
2026 = '251.94'
2027 = '261.05'
2028 = '270.17'
2029 = '279.28'
2030 = '288.40'
2031 = '299.22'
2032 = '310.05'
2033 = '320.87'
```

```
  2034 = '331.70'
  2035 = '342.52'
  2036 = '355.38'
  2037 = '368.24'
  2038 = '381.09'
  2039 = '393.95'
  2040 = '406.81'
  2041 = '422.08'
  2042 = '437.35'
  2043 = '452.62'
  2044 = '467.89'
  2045 = '483.16'
  2046 = '501.30'
  2047 = '519.44'
  2048 = '537.57'
  2049 = '555.71'
  2050 = '573.85'
  2051 = '595.39'
  2052 = '616.93'
  2053 = '638.47'
  2054 = '660.01'
  2055 = '681.55'
  2056 = '707.13'
  2057 = '732.72'
  2058 = '758.30'
  2059 = '783.89'
  2060 = '809.47'
  2061 = '839.85'
  2062 = '870.24'
  2063 = '900.62'
  2064 = '931.01'
  2065 = '961.39'
  2066 = '997.48'
  2067 = '1033.57'
  2068 = '1069.66'
  2069 = '1105.75'
  2070 = '1141.84'
  2071 = '1184.70'
  2072 = '1227.56'
  2073 = '1270.42'
  2074 = '1313.28'
  2075 = '1356.14'
  other = 'ERROR'
;
run;
```

B.2. Source Code of ssawage.sas

This appendix lists the source code of ssawage.sas This file defines a format to map calendar years into the corresponding (projected) Social Security average wage index. Future values are based on intermediate assumptions of assumptions of OASDI Board of Trustees (1998).

Source: OASDI Board of Trustees (1998), Table III.B1. Projections for individual years after 2007 were interpolated linearly.

Source Code

```
proc format;
value ssawage
  1951 = '2799.16'
  1952 = '2973.32'
  1953 = '3139.44'
  1954 = '3155.64'
  1955 = '3301.44'
  1956 = '3532.36'
  1957 = '3641.72'
  1958 = '3673.80'
  1959 = '3855.80'
  1960 = '4007.12'
  1961 = '4086.76'
  1962 = '4291.40'
  1963 = '4396.64'
  1964 = '4576.32'
  1965 = '4658.72'
  1966 = '4938.36'
  1967 = '5213.44'
  1968 = '5571.76'
  1969 = '5893.76'
  1970 = '6186.24'
  1971 = '6497.08'
  1972 = '7133.80'
  1973 = '7580.16'
  1974 = '8030.76'
  1975 = '8630.92'
  1976 = '9226.48'
  1977 = '9779.44'
  1978 = '10556.03'
  1979 = '11479.46'
  1980 = '12513.46'
  1981 = '13773.10'
  1982 = '14531.34'
  1983 = '15239.24'
  1984 = '16135.07'
```

```
1985 = '16822.51'
1986 = '17321.82'
1987 = '18426.51'
1988 = '19334.04'
1989 = '20099.55'
1990 = '21027.98'
1991 = '21811.60'
1992 = '22935.42'
1993 = '23132.67'
1994 = '23753.53'
1995 = '24705.66'
1996 = '25913.90'
1997 = '27019.16'
1998 = '27894.53'
1999 = '28835.56'
2000 = '29919.10'
2001 = '30988.90'
2002 = '32128.11'
2003 = '33428.05'
2004 = '34870.86'
2005 = '36380.25'
2006 = '37952.99'
2007 = '39613.01'
2008 = '41433.80'
2009 = '43254.59'
2010 = '45075.38'
2011 = '47241.08'
2012 = '49406.78'
2013 = '51572.47'
2014 = '53738.17'
2015 = '55903.87'
2016 = '58589.84'
2017 = '61275.81'
2018 = '63961.77'
2019 = '66647.74'
2020 = '69333.71'
2021 = '72664.93'
2022 = '75996.15'
2023 = '79327.36'
2024 = '82658.58'
2025 = '85989.80'
2026 = '90121.27'
2027 = '94252.75'
2028 = '98384.23'
2029 = '102515.71'
2030 = '106647.19'
2031 = '111771.17'
2032 = '116895.16'
2033 = '122019.14'
2034 = '127143.12'
2035 = '132267.11'
2036 = '138622.03'
```

```
  2037 = '144976.95'
  2038 = '151331.89'
  2039 = '157686.81'
  2040 = '164041.73'
  2041 = '171923.31'
  2042 = '179804.88'
  2043 = '187686.45'
  2044 = '195568.02'
  2045 = '203449.59'
  2046 = '213224.56'
  2047 = '222999.53'
  2048 = '232774.52'
  2049 = '242549.48'
  2050 = '252324.45'
  2051 = '264447.69'
  2052 = '276570.91'
  2053 = '288694.12'
  2054 = '300817.34'
  2055 = '312940.56'
  2056 = '327976.16'
  2057 = '343011.72'
  2058 = '358047.31'
  2059 = '373082.88'
  2060 = '388118.47'
  2061 = '406766.06'
  2062 = '425413.66'
  2063 = '444061.28'
  2064 = '462708.88'
  2065 = '481356.47'
  2066 = '504483.81'
  2067 = '527611.12'
  2068 = '550738.44'
  2069 = '573865.81'
  2070 = '596993.12'
  2071 = '625676.31'
  2072 = '654359.56'
  2073 = '683042.75'
  2074 = '711726.00'
  2075 = '740409.19'
  other = 'ERROR'
;
run;
```

B.3. Source Code of `marstat.sas`

This appendix lists the source code of `marstat.sas` This file defines a macro to determine an individual’ s (projected) marital status as of a certain date. Historical values correspond to information in the Survey of Income and Program Participation (SIPP); future values follow from MINT’ s demographic projections. For married individuals, the macro also returns the marriage order number.

The line numbers on the far left of the listing are not part of the code. They are included to ease description of the code. They do not appear in the actual source code.

Source Code

```
/*************************************************************************/
/*                                                                       */
/*  Define macro to figure out demographic status as of                */
/*  a certain date.                                                      */
/*                                                                       */
/*  Arguments:                                                           */
/*       date   [in] SAS date (days since 1/1/1960)                      */
/*       status [out] marital status as of date:                         */
/*                    0 = never married                                  */
/*                    1= married                                        */
/*                    2 = widowed                                        */
/*                    3 = divorced                                       */
/*                    4 = deceased                                     */
/*       marnum [out] marriage number (if status=1, 2, or 3)             */
/*                                                                       */
/*  Requires that the following variables and arrays have                */
/*  been defined in the data step:                                       */
/*       nummar            # marriages until death                       */
/*       deathdte          Date of death                                 */
/*       howend1-howend12  How did marriage end?                        */
/*       marb_1-marb_12    Wedding date                                  */
/*       mare_1-mare_12    Marriage dissolution date                     */
/*       spbdat1-spbdat12  Spousal birth date                          */
/*                                                                       */
/*  Note: the following arrays must have been defined in the             */
/*  calling program:                                                   / *
/*     array howend(*)  howend1-howend12;                                */
/*     array marb(*)    marb_1-marb_12;                                  */
/*     array mare(*)    mare_1-mare_12;                                  */
/*     array spbdate(*) spbdat1-spbdat12;                               */
/*                                                                       */
/*  Note: this macro is similar to %figstat used elsewhere in the        */
/*  MINT project.  It differes in that %marstat returns the marriage  */
/*  number.                                                              */
/*                                                                       */
/*  Stan Panis, 2 August 1999                                         */
/*                                                                       */
/*************************************************************************/

```

```
%macro marstat(date,status,marnum);
   &status=.;
   &marnum=.;
   if (deathdte^=. and &date>=deathdte) then &status=4;  /* deceased */
   else if (nummar=0) then &status=0; /* never married */
   else if (&date<marb(1)) then &status=0;  /* never married */
   else do;
      do ii=1 to nummar while (&status=.);
         if (marb(ii)<=&date<mare(ii)) then do;
            &status=1;  /* married */
         end; else if (&date<marb(ii)) then do;
            if (howend(ii-1)=0) then &status=1;      /* married */
            else if (howend(ii-1)=1) then &status=3; /* divorced */
            else if (howend(ii-1)=2) then &status=2; /* widowed */
            else if (howend(ii-1)=3) then &status=4; /* deceased */
            else put "Error 1 in program logic!";
         end; else if (ii=nummar and &date>=mare(nummar)) then do;
            if (howend(ii)=0) then &status=1;      /* married */
            else if (howend(ii)=1) then &status=3; /* divorced */
            else if (howend(ii)=2) then &status=2; /* widowed */
            else if (howend(ii)=3) then &status=4; /* deceased */
            else put "Error 2 in program logic!";
         end;
         if (&status=1 or &status=2 or &status=3) then &marnum=ii;
      end;
   end;
   if (&status=.) then put "Error 3 in program logic!";
   drop ii;
%mend;
```

B.4. Source Code of **`computax.sas`**

This appendix lists the source code of `computax`, the main macro to compute MINT tax liabilities.

The line numbers on the far left of the listing are not part of the code. They are included to ease description of the code. They do not appear in the actual source code.

Source Code

```
%macro computax(year,assetinc,fedtax,ficatax,statetax);

   /*  Macro to compute or approximate federal income tax, Federal    */
   /*  Insurance Contributions Act (FICA), and state total tax        */
   /*  liability.  Arguments:                                         */
   /*                                                                 */
   /*     year     = fiscal year, e.g., 2020                          */
   /*     assetinc = array with asset income flows, e.g., inci        */
   /*     fedtax   = federal income taxes owed                        */
   /*     statetax = state total taxes owed                           */
   /*     ficatax  = Federal Insurance Contributions Act taxes owed   */
   /*                                                                 */
   /*  This macro is extensively documented in User's Guide to SSA's  */
   /*  MINT Tax Model, Klerman and Panis, August 1999.                */
   /*                                                                 */
   /*  Jacob Klerman and Stan Panis, 11 August 1999                   */

   /* State total tax liability is lambda times  */
   /* federal income tax liability.              */
   %let lambda=0.835;

   /* Fraction of annuitized wealth (asset income) that is taxable */
   %let gamma=1;

   length cpi ssawage 4;
   format &fedtax &ficatax &statetax 8.2;

   /* Check the validity of the year */
   if (&year<1990 or &year>2031) then do;
      put "ERROR: year input to computax must be between 1990 and 2031.";
      abort;
   end;

   /* Compute some basic variables:                                    */
   /*    married = indicator for married as of January 1 of next year  */
   /*              Note: recently widowed may file as married.         */
   /*    age     = age as of December 31                               */
   /*    spage   = spousal age, if any, as of December 31              */
   /*    cpi     = Consumer Price Index for fiscal year                */
   /*    ssawage = Social Security average earnings for fiscal year    */
   tmp = mdy(1,1,&year+1);
   %marstat(tmp,married,marnum);
```

```
if (married=2) then do;  /* widowed */
   /* If widowed in reference year, survivor may file as married */
   if (year(mare(marnum))=&year) then married=1;
end;
married = (married=1);
age = &year-year(brthdate);  /* age as of end of year */
if (married=1) then spage = &year - year(spbdate(marnum));


cpi = put(&year,cpi.)/put(1998,cpi.);
ssawage = put(&year,ssawage.);

/*  Four classes of income (all in current dollars)          */
/*     cei = couple's earned income                          */
/*     cdb = couple's defined benefit pension income         */
/*     cra = couple's taxable return on assets               */
/*     css = couple's Social Security benefits               */
/*  Note: imputed rental income is not subject to taxation.  */

/*  earned income (keep track of both spouses because of FICA)  */
if (married=0) then do;
   hei = inde(&year) * ssawage;
   cei = hei;
end; else do;
   hei = inde(&year) * ssawage;
   sei = sern(&year) * ssawage;
   cei = hei+sei;
end;

/*  defined benefit pension payments */
if (married=0) then cdb = hpen(&year) * ssawage;
else cdb = (hpen(&year) + spen(&year)) * ssawage;

/*  return on financial assets (fraction gamma assumed taxable) */
cra = &gamma * &assetinc(&year)*ssawage;

/*  taxable social security income */
if (married=0) then css = ssb(&year) * ssawage;
else css = (ssb(&year) + sssb(&year)) * ssawage;
/*  SS benefit taxability rules                                 */
/*  For SS benefit taxability test, income is defined as AGI      */
/*  w/o SS benefits plus half SS benefits.                        */
definc=cei+cdb+cra+0.5*css;
/*  Code below follows worksheet on page 27 of the 1998 Form 1040A */
/*  instructions.  N.B.:  These cutoffs are NOT indexed.          */
if (married=0) then do; W8=25000; W10=9000; end;
else do; W8=32000; W10=12000; end;
W9 =max(definc-W8, 0);
W11=max(W9-W10, 0);
W12=min(W9, W10);
W14=min(0.5*css, 0.5*W12);
W15=max(0.85*W11, 0);
acss=min(W14+W15, 0.85*css);

/****************************************************************/
/*  Federal income tax                                         */
/*  Variable names indicate line number on the 1998 Form 1040A  */
```

```
/*****************************************************************/

A18 = cei+cdb+acss+cra;   /* Adjusted Gross Income */

/* Standard deduction */
if (married=0) then do;
   A20a = (age>=65);
   if (A20a=0) then A21 = 4250*cpi;
   else if (A20a=1) then A21 = 5300*cpi;
end; else do;
   A20a = (age>=65) + (spage>=65);
   if (A20a=0) then A21 = 7100*cpi;
   else if (A20a=1) then A21 = 7950*cpi;
   else if (A20a=2) then A21 = 8800*cpi;
end;
A22 = max(0, A18-A21);

/* Personal exemptions */
if (married=0) then A23 = 1*2700*cpi;
else A23 = 2*2700*cpi;
A24 = max(0, A22-A23);

/* federal income tax (marginal and incremental marginal rates) */
ftr1=0.150;
ftr2=0.280;
ftr3=0.310;
ftr4=0.360;
ftr5=0.396;
fitr1=ftr1;
fitr2=ftr2-ftr1;
fitr3=ftr3-ftr2;
fitr4=ftr4-ftr3;
fitr5=ftr5-ftr4;

if (married=0) then do;
   ftbrac1= 25350*cpi;
   ftbrac2= 61400*cpi;
   ftbrac3=128100*cpi;
   ftbrac4=278450*cpi;
end; else do;
   ftbrac1= 42350*cpi;
   ftbrac2=102300*cpi;
   ftbrac3=155950*cpi;
   ftbrac4=278450*cpi;
end;

/* Federal income tax */
&fedtax = fitr1*max(0,A24)+
          fitr2*max(0,A24-ftbrac1)+
          fitr3*max(0,A24-ftbrac2)+
          fitr4*max(0,A24-ftbrac3)+
          fitr5*max(0,A24-ftbrac4);

/* Tax credit for the elderly and disabled (1998 Form 1040  */
/* Schedule 3).  MINT excludes DI so only tax credit for    */
/* elderly is accounted for.                                */
C10=0;
```

```
   C15=0;
   if (married=0) then do;
      if (age>=65) then do;
         C10=5000*cpi;
         C15=7500*cpi;
      end;
   end; else do;
      tmp = (age>=65) + (spage>=65);
      if (tmp=1) then do;
         C10= 5000*cpi;
         C15=10000*cpi;
      end; else if (tmp=2) then do;
         C10= 7500*cpi;
         C15=10000*cpi;
      end;
   end;
   C13a=max(0,css-acss);
   C16=max(0,A18-C15);
   C18=C13a+0.5*C16;
   A27=0.15*max(0,C10-C18);
   &fedtax = &fedtax - A27;

   /* Federal Insurance Contributions Act (FICA) tax.                */
   /* OASDI capped at $68,400 in 1998, indexed by SSA average wage. */
   ficacap = 68400*put(&year,ssawage.)/put(1998,ssawage.);
   if (married=0) then &ficatax = 0.0620*min(hei, ficacap) + 0.0145*hei;
   else &ficatax = 0.0620*min(hei, ficacap) +
                   0.0620*min(sei, ficacap) +
                   0.0145*cei;

   /* Estimated state total tax */
   &statetax = &lambda*fedtax;

   drop married age spage cpi ssawage hei sei cei cdb cra css acss
        marnum definc W8 W10 W9 W11 W12 W14 W15 A18 A20a A21
        A22 A23 A24 A27 C10 C15 C13a C16 C18 tmp
        ftr1 fitr1 ftr2 fitr2 ftr3 fitr3 ftr4 fitr4
        ftr5 fitr5 ftbrac1 ftbrac2 ftbrac3 ftbrac4 ficacap;
%mend;
```

B.5. 1998 U.S. Individual Income Tax Return Form 1040A

Form **1040A** Department of the Treasury—Internal Revenue Service

U.S. Individual Income Tax Return **1998** IRS Use Only—Do not write or staple in this space.

OMB No. 1545-0085

Label (See page 18.)

Use the IRS label. Otherwise, please print or type.

LABEL HERE

Your first name and initial | Last name | **Your social security number**

If a joint return, spouse's first name and initial | Last name | **Spouse's social security number**

Home address (number and street). If you have a P.O. box, see page 19. | Apt. no.

City, town or post office, state, and ZIP code. If you have a foreign address, see page 19.

▲ IMPORTANT! ▲ You **must** enter your SSN(s) above.

Presidential Election Campaign Fund (See page 19.) Yes No

Do you want $3 to go to this fund?

If a joint return, does your spouse want $3 to go to this fund?

Note: *Checking "Yes" will not change your tax or reduce your refund.*

Filing status

Check only one box.

1 ☐ Single

2 ☐ Married filing joint return (even if only one had income)

3 ☐ Married filing separate return. Enter spouse's social security number above and full name here. ▶

4 ☐ Head of household (with qualifying person). (See page 20.) If the qualifying person is a child but not your dependent, enter this child's name here. ▶

5 ☐ Qualifying widow(er) with dependent child (year spouse died ▶ 19). (See page 21.)

Exemptions

6a ☐ **Yourself.** If your parent (or someone else) can claim you as a dependent on his or her tax return, **do not** check box 6a.

b ☐ **Spouse**

No. of boxes checked on 6a and 6b

c **Dependents:**

If more than seven dependents, see page 21.

(1) First name Last name	(2) Dependent's social security number	(3) Dependent's relationship to you	(4) ✓ if qualified child for child tax credit (see page 22)
			☐
			☐
			☐
			☐
			☐
			☐
			☐

No. of your children on 6c who:

- **lived with you**
- **did not live with you due to divorce or separation (see page 23)**

Dependents on 6c not entered above

d Total number of exemptions claimed.

Add numbers entered on lines above

Income

Attach Copy B of your Forms W-2 and 1099-R here.

If you did not get a W-2, see page 24.

Enclose, but do not staple, any payment.

7 Wages, salaries, tips, etc. Attach Form(s) W-2. 7

8a **Taxable** interest. Attach Schedule 1 if required. 8a

b **Tax-exempt** interest. DO NOT include on line 8a. 8b

9 Ordinary dividends. Attach Schedule 1 if required. 9

10a Total IRA distributions. 10a | 10b Taxable amount (see page 24). 10b

11a Total pensions and annuities. 11a | 11b Taxable amount (see page 25). 11b

12 Unemployment compensation. 12

13a Social security benefits. 13a | 13b Taxable amount (see page 27). 13b

14 Add lines 7 through 13b (far right column). This is your **total income.** ▶ 14

Adjusted gross income

15 IRA deduction (see page 28). 15

16 Student loan interest deduction (see page 28). 16

17 Add lines 15 and 16. These are your **total adjustments.** 17

18 Subtract line 17 from line 14. This is your **adjusted gross income.** If under $30,095 (under $10,030 if a child did not live with you), see the EIC instructions on page 36. ▶ 18

For Disclosure, Privacy Act, and Paperwork Reduction Act Notice, see page 49. Cat. No. 11327A **1998 Form 1040A**

1998 Form 1040A page 2

Taxable income

19 Enter the amount from line 18. 19

20a Check if: ☐ **You** were 65 or older ☐ Blind ☐ **Spouse** was 65 or older ☐ Blind } **Enter number of boxes checked** ▶ 20a

b If you are married filing separately and your spouse itemizes deductions, see page 30 and check here . . . . ▶ 20b ☐

21 Enter the **standard deduction** for your filing status. **But** see page 31 if you checked any box on line 20a or 20b **OR** if someone can claim you as a dependent.
- Single—$4,250
- Married filing jointly or Qualifying widow(er)—$7,100
- Head of household—$6,250
- Married filing separately—$3,550

21

22 Subtract line 21 from line 19. If line 21 is more than line 19, enter -0-. 22

23 Multiply $2,700 by the total number of exemptions claimed on line 6d. 23

24 Subtract line 23 from line 22. If line 23 is more than line 22, enter -0-. This is your **taxable income.** ▶ 24

Tax, credits, and payments

25 Find the tax on the amount on line 24 (see page 31). 25

26 Credit for child and dependent care expenses. Attach Schedule 2. 26

27 Credit for the elderly or the disabled. Attach Schedule 3. 27

28 Child tax credit (see page 32). 28

29 Education credits. Attach Form 8863. 29

30 Adoption credit. Attach Form 8839. 30

31 Add lines 26 through 30. These are your **total credits.** 31

32 Subtract line 31 from line 25. If line 31 is more than line 25, enter -0-. 32

33 Advance earned income credit payments from Form(s) W-2. 33

34 Add lines 32 and 33. This is your **total tax.** ▶ 34

35 Total Federal income tax withheld from Forms W-2 and 1099. 35

36 1998 estimated tax payments and amount applied from 1997 return. 36

37a **Earned income credit.** Attach Schedule EIC if you have a qualifying child. 37a

b Nontaxable earned income: amount ▶ and type ▶

38 Additional child tax credit. Attach Form 8812. 38

39 Add lines 35, 36, 37a, and 38. These are your **total payments.** ▶ 39

Refund

Have it directly deposited! See page 43 and fill in 41b, 41c, and 41d.

40 If line 39 is more than line 34, subtract line 34 from line 39. This is the amount you **overpaid.** 40

41a Amount of line 40 you want **refunded to you.** 41a

b Routing number

c Type: ☐ Checking ☐ Savings

d Account number

42 Amount of line 40 you want **applied to your 1999 estimated tax.** 42

Amount you owe

43 If line 34 is more than line 39, subtract line 39 from line 34. This is the **amount you owe.** For details on how to pay, see page 44. 43

44 Estimated tax penalty (see page 44). 44

Sign here

Joint return? See page 19.

Keep a copy for your records.

Under penalties of perjury, I declare that I have examined this return and accompanying schedules and statements, and to the best of my knowledge and belief, they are true, correct, and accurately list all amounts and sources of income I received during the tax year. Declaration of preparer (other than the taxpayer) is based on all information of which the preparer has any knowledge.

Your signature	Date	Your occupation	Daytime telephone number (optional)
Spouse's signature. If joint return, BOTH must sign.	Date	Spouse's occupation	( )

Paid preparer's use only

Preparer's signature	Date	Check if self-employed ☐	Preparer's social security no.
Firm's name (or yours if self-employed) and address			EIN
			ZIP code

B.6. 1998 Form 1040A—Social Security Benefits Worksheet

Social Security Benefits Worksheet—Lines 13a and 13b (keep for your records)

If you are married filing separately and you **lived apart** from your spouse for all of 1998, enter "D" in the space to the right of the word "benefits" on line 13a.

1. Enter the total amount from **box 5** of **all** your **Forms SSA-1099** and **Forms RRB-1099** 1. ______

Note: *If line 1 is zero or less,* ***stop****; none of your social security benefits are taxable. Otherwise, go to line 2.*

2. Enter one-half of line 1 . . . 2. ______

3. Add the amounts on Form 1040A, lines 7, 8a, 9, 10b, 11b, and 12. Do not include amounts from box 5 of Forms SSA-1099 or RRB-1099 . . . 3. ______

4. Enter the amount, if any, from Form 1040A, line 8b . . . 4. ______

5. Add lines 2, 3, and 4 . . . 5. ______

6. Enter the amount, if any, from Form 1040A, line 15 . . . 6. ______

7. Subtract line 6 from line 5 . . . 7. ______

8. Enter: $25,000 if single, head of household, qualifying widow(er), or married filing separately and you **lived apart** from your spouse for all of 1998; $32,000 if married filing jointly; -0- if married filing separately and you lived with your spouse at any time during 1998 . . . 8. ______

9. Subtract line 8 from line 7. If zero or less, enter -0- . . . 9. ______

Is line 9 more than zero?

No. Stop; none of your social security benefits are taxable. You do not have to enter any amount on line 13a or 13b of Form 1040A. **But** if you are married filing separately and you **lived apart** from your spouse for all of 1998, enter -0- on line 13b. Be sure to enter "D" to the right of the word "benefits" on line 13a.

Yes. Go to line 10.

10. Enter: $9,000 if single, head of household, qualifying widow(er), or married filing separately and you **lived apart** from your spouse for all of 1998; $12,000 if married filing jointly; -0- if married filing separately and you lived with your spouse at any time during 1998 . . . 10. ______

11. Subtract line 10 from line 9. If zero or less, enter -0- . . . 11. ______

12. Enter the **smaller** of line 9 or line 10 . . . 12. ______

13. Enter one-half of line 12 . . . 13. ______

14. Enter the **smaller** of line 2 or line 13 . . . 14. ______

15. Multiply line 11 by 85% (.85). If line 11 is zero, enter -0- . . . 15. ______

16. Add lines 14 and 15 . . . 16. ______

17. Multiply line 1 by 85% (.85) . . . 17. ______

18. Taxable social security benefits. Enter the **smaller** of line 16 or line 17 . . . 18. ______

- Enter the amount from line 1 on Form 1040A, line 13a.
- Enter the amount from line 18 on Form 1040A, line 13b.

TIP *If part of your benefits are taxable for 1998* ***and*** *they include benefits paid in 1998 that were for an earlier year, you may be able to reduce the taxable amount shown on the worksheet. See Pub. 915 for details.*

B.7. 1998 Form 1040A—Schedule 3

Schedule 3 (Form 1040A)

Department of the Treasury—Internal Revenue Service

Credit for the Elderly or the Disabled for Form 1040A Filers (99) **1998**

OMB No. 1545-0085

Name(s) shown on Form 1040A | **Your social security number**

You may be able to take this credit and reduce your tax if by the end of 1998:

- You were age 65 or older, **OR**
- You were under age 65, you retired on **permanent and total** disability, and you received taxable disability income.

But you must also meet other tests. See the separate instructions for Schedule 3.

TIP: In most cases, the IRS can figure the credit for you. See the instructions.

Part I

Check the box for your filing status and age

If your filing status is:	**And by the end of 1998:**	**Check only one box:**
Single, Head of household, or Qualifying widow(er) with dependent child	**1** You were 65 or older	1 ☐
	2 You were under 65 and you retired on permanent and total disability	2 ☐
Married filing a joint return	**3** Both spouses were 65 or older	3 ☐
	4 Both spouses were under 65, but only one spouse retired on permanent and total disability	4 ☐
	5 Both spouses were under 65, and both retired on permanent and total disability	5 ☐
	6 One spouse was 65 or older, and the other spouse was under 65 and retired on permanent and total disability	6 ☐
	7 One spouse was 65 or older, and the other spouse was under 65 and **NOT** retired on permanent and total disability	7 ☐
Married filing a separate return	**8** You were 65 or older and you lived apart from your spouse for all of 1998	8 ☐
	9 You were under 65, you retired on permanent and total disability, and you lived apart from your spouse for all of 1998	9 ☐

Did you check box 1, 3, 7, or 8?

Yes ▶ Skip Part II and complete Part III on the back.

No ▶ Complete Parts II and III.

Part II

Statement of permanent and total disability

Complete this part **only** if you checked box 2, 4, 5, 6, or 9 above.

IF: **1** You filed a physician's statement for this disability for 1983 or an earlier year, or you filed a statement for tax years after 1983 and your physician signed line B on the statement, **AND**

2 Due to your continued disabled condition, you were unable to engage in any substantial gainful activity in 1998, check this box . . . . ▶ ☐

- If you checked this box, you do not have to get another statement for 1998.
- If you **did not** check this box, have your physician complete the statement on page 4 of the instructions. You **must** keep the statement for your records.

For Paperwork Reduction Act Notice, see Form 1040A instructions. Cat. No. 12064K **1998 Schedule 3 (Form 1040A)**

Part III
Figure your credit

10 **If you checked (in Part I):** **Enter:**

Box 1, 2, 4, or 7	$5,000	
Box 3, 5, or 6	$7,500	
Box 8 or 9	$3,750	10

Did you check box 2, 4, 5, 6, or 9 in Part I?

- **Yes** → You **must** complete line 11.
- **No** → Enter the amount from line 10 on line 12 and go to line 13.

11
- If you checked box 6 in Part I, add $5,000 to the taxable disability income of the spouse who was under age 65. Enter the total.
- If you checked box 2, 4, or 9 in Part I, enter your taxable disability income.
- If you checked box 5 in Part I, add your taxable disability income to your spouse's taxable disability income. Enter the total.

TIP: For more details on what to include on line 11, see the instructions. 11

12 If you completed line 11, enter the **smaller** of line 10 or line 11; **all others,** enter the amount from line 10. 12

13 Enter the following pensions, annuities, or disability income that you (and your spouse if filing a joint return) received in 1998.

a Nontaxable part of social security benefits, and

Nontaxable part of railroad retirement benefits treated as social security. See instructions. 13a

b Nontaxable veterans' pensions and any other pension, annuity, or disability benefit that is excluded from income under any other provision of law. See instructions. 13b

c Add lines 13a and 13b. (Even though these income items are not taxable, they **must** be included here to figure your credit.) If you did not receive any of the types of nontaxable income listed on line 13a or 13b, enter -0- on line 13c. 13c

14 Enter the amount from Form 1040A, line 19. 14

15 **If you checked (in Part I):** **Enter:**

Box 1 or 2	$7,500	
Box 3, 4, 5, 6, or 7	$10,000	
Box 8 or 9	$5,000	15

16 Subtract line 15 from line 14. If zero or less, enter -0-. 16

17 Enter one-half of line 16. 17

18 Add lines 13c and 17. 18

19 Subtract line 18 from line 12. If zero or less, **stop;** you **cannot** take the credit. Otherwise, go to line 20. 19

20 Multiply line 19 by 15% (.15). Enter the result here and on Form 1040A, line 27. 20